Bioinformatic analysis of characteristics and structure of the human genome: tracking the footprints of natural selection mediated by gene expression

Odabachian, Araxi Urrutia

Award date:
2003

Awarding institution:
University of Bath

Bioinformatic Analysis of Characteristics and Structure of the Human Genome: Tracking the Footprints of Natural Selection Mediated by Gene Expression

AUTHOR: A (paxi) URRUTIA ODABACHIAN **YEAR: 2003**

TITLE : BIOINFORMATIC ANALYSIS OF CHARACTERISTICS AND STRUCTURE OF THE HUMAN GENOME: TRACKING THE FOOTPRINTS OF NATURAL SELECTION MEDIATED BY GENE EXPRESSION

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Department of Biology and Biochemistry

University of Bath

Supervisor: Prof. Laurence D. Hurst

Assessor: Prof. Richard ffrench-Constant

External examiner: Prof. Brian Charlesworth

To Professor Laurence Hurst

Summary

During my PhD I analysed whether our genes and genome are adapted for optimization of protein synthesis. At the time I started my graduate studies the common thought was that, in humans, gene characteristics were determined almost exclusively by mutational processes on the one hand, and the optimization of function of the encoded protein on the other. From the results obtained so far during my PhD, we are confronted with a different view, one in which genome structure becomes an active and important player in the regulation of large scale patterns of gene activity, and genes are not only under pressure to produce proteins with optimum enzymatic function but also for doing it cheaply.

Initially I studied codon usage bias (CUB), that is, the unequal use of codons that encode for the same amino acid. In other species, CUB is closely related to expression patterns: genes of greater expression are more biased. In mammals, however, no such relation had been established. The human genome is highly heterogeneous in its base composition, which had hampered previous efforts to examine patterns of codon usage. I developed a new method to calculate CUB that could account for background nucleotide bias. Contrary to previous expectations, I found significant biases in codon usage in human genes not explainable by background nucleotide distributions.

With the publication of the human genome draft sequence and the greater availability of gene expression data, I was able to show that highly expressed genes encode shorter proteins, have smaller introns, and have greater biases in their codon distributions and amino acid use. All these patterns are precisely what we would expect if selective pressures act to reduce costs of protein synthesis.

Previously, genes were generally believed to be randomly located in the genome. I found that genes with greater expression, on average, are to be found near to other genes of broad expression. This observation was among the first to describe a general pattern of gene sorting along the genome. Moreover, I found a relationship between the tight coupling of gene expression and chromosome location with the puzzling base composition heterogeneity (isochore structures). I observed that the base composition of flanking non-transcribed regions of the genes is greatly correlated with their expression levels. The higher the G+C content, the higher the expression.

Acknowledgements

I am very fortunate to have received support and friendship of many people who have made it possible for me to be so far from home and get to finally finish writing this thesis. I wish to thank all the people who have shared with me many happy moments and to those who had lent a hand when I most needed it.

My sincere gratitude to my supervisor Prof. Laurence Hurst for the opportunity to study in one of the best groups in the field of molecular evolution, but most importantly for never giving up on me.

To my family who have kept an eye on me during this time, muchas gracias, I especially thank my mother for being there for me in the phone and visiting me several times. No more 10-hour-long flights Ma! Muchas gracias. I also thank my father Jaime Urrutia, my brother Avedis Urrutia for coming all the way to visit me. I also thank my grandmother Margarita Fucugauchi for accompanying me all this time through her letters and later through emails. Me siento muy afortunada de que sean mi familia.

Sections of this thesis contain work done in collaboration with Dr. Martin Lercher and Dr. Adam Pavlíček. It has been a pleasure to work with them.

I wish to thank Prof. Brian Charlesworth for his advice and comments, even before the start of my program, and for acting as the external examiner. Thanks to Dr. Alan Wheals for all his advice throughout the program, also for his support in getting funds to attend to conferences. I thank Prof. Richard ffrench-Constant for his comments and his role as advisor and internal examiner.

Thanks to all in the department who helped me with the all the everyday matters. In particular I want to thank Amanda Harper, Alan King, Anne Pearson, Helen Parkin and all the people at the finance office for all their help. Sorry for troubling you so much with last minute emergencies. I also thank all the security people who helped every time I got locked out at night, many thanks. I apologise for never putting down my name in the after-hours list.

I thank Dr. Paul Allen, Dr. Matthew Wills, Dr. Alan Rayner and all who gave me the chance to demonstrate at their lab practicals, I learnt a lot, I hope the students did too.

I had the great opportunity to visit the labs of people at other universities where I was always welcomed in the best of ways. These visits were very enriching. I want to thank Prof. Wen Hsiung Li from Chicago University, Prof. Blair Hedges from Penn State University, Dr. Alexey Kondrashov from the NCBI, Dr. Sudhir Kumar from Arizona State University. I wish to express my gratitude to Dr. Hiroshi Akashi at Penn State University for all his help and for allowing me to stay as a visitor student at his friendly inspiring high tech lab. Also I thank his students Anusha Radakrishnan, Wen Ya Ko and Anoop John for all their help and friendship.

I would also like to thank those who played an important role in my arrival in Bath; my especial thanks go to Dr. Federico Bermudez, Dr. Victor Ramirez, Dr. Arturo Bouzas and Dr. Dante Moran Centeno, from the National University of Mexico and to Dr. Kenneth Hastings, Dr. Angel Alonso and Dr. Ron Chase at McGill University in Canada.

Gracias a Gustavo Ortiz por ser un amigo. Gracias a Liliana Beltran por ser mi amiga. Elizabeth Williams, the best labmate I had. Very good times chatting for hours at the lab. I'm in debt to Gabriela Roca, José Fernando Mikan, Kim Reilly, Lilian Madi-

Ravazzi, Shalini Iyer, Paschalis Giannoulis, Alessandra Spila, Paul Dean, Adao Dos Santos, R. Natesh, José Luis Morales, Ana Herrera, Claudia Carrillo, Martín Alejandro Serrano, Adedapo Gbadegesin, Mikki Koo, Csaba Pal, Jean-Vincent Chamary, and all those who are or were at the department and made the difference to my time at the *south building.* Also thanks to Gerardo Cuervo Chargoy, Narendra Buch, Miguel Lara, Martha Lopez Araiza and Carlos Arango. Gracias a todos.

My thanks to Jaleh and Barry d'Archambaud for opening the doors of their home for me and having me as tenant for three years.

Gracias a Humberto Gutiérrez for all his help during these years, for discussing and proof-reading several sections of this thesis, y por estar siempre cerca.

I want to thank to the institutions that provided me with the funds to support myself during my PhD. I specially thank CONACyT for the scholarship that covered my living expenses and tuition which actually made possible my arrival to the UK. I also want to thank the ORS for the award to partly cover my tuition fees during the last two years of my PhD; the Department of Biology and Biochemistry at the University of Bath. Thanks to a Korner award from Sussex University as well as further support from the host university I was able to visit Dr. Hiroshi Akashi's lab at Penn State University. In addition, I received additional support to cover the costs of attending several conferences. These funds were provided by the Dept of Biology and Biochemistry at Bath University, the American Genetics Association and the Genetics Society. Many thanks to all those involved. Finally I want to thank Dr. Sudhir Kumar, Director of the Evolution and Functional Genomics Center at Arizona State University for funding the travel expenses to fly back to the UK to present the viva.

Work Description

The thesis work is composed of four papers. In all of them I got involved to different degrees on all stages from the setting of the project to the manuscript writing and the reply to referees. My particular contributions for the different chapters are as follows. Chapter 2. I did not set the initial project. I developed the method to measure codon bias and the methodology to correct for background nucleotide biases. I designed all tests except for the use of randomizations. I performed all analyses included in the chapter, except for those presented in figure 1 which showws the relationship between GC content in coding and non coding regions. I wrote the initial draft of the manuscript and various later versions. All versions where extensively edited by my supervisor to obtain the final version. Chapter 3. I developed the initial of layout the project. I designed all analyses included. I performed all analyses there included. I wrote the initial draft of the manuscript and various later versions. All versions where extensively edited by my supervisor to obtain the final version. Chapter 4. I developed the initial of layout the project. I designed the analyses regarding the heterogeneity in expression levels of different chromosomes. I designed analyses regarding the correlations among the various measures of expression measures and to different gene characteristics and their context. I designed an earlier version of the analysis regarding local similarity in expression profiles (but not included in the final version of the manuscript). I performed analyses regarding the heterogeneity of average expression of genes according to chromosome. I performed analyses relationships between different expression measures and gene characteristics and genetic context. I performed analyses on the local similarity of expression profiles of genes (not included in the final version). I reviewed the manuscript prior to publication. Chapter 5. I did not set the layout

of the project. I suggested the analyses of the impact of recombination on the variables under study. I performed confirmatory analyses using an independent dataset of expression profiles. I wrote the initial draft of the manuscript and various later versions. All versions where extensively edited by co authors of the work and my supervisor to obtain the final version.

CONTENTS

Chapter One

General Introduction

GENERAL INTRODUCTION

The human genome sequencing project constitutes one of the greatest collaborative efforts in biology. This enormous enterprise has been fuelled by the expectation that the sequence will provide answers about function of genes, the nature of inherited diseases, and suggest suitable targets for drug development. After 20 years, while important progress has been achieved, the poor understanding we still have of the function of most genes is evident. Even less we know about the ways in which genes coordinate their expression during development, maintain body functions and allow us to respond to environmental challenges. With the completed human genome available, it seems clear that reading *the book of life* will be trickier than first expected.

The genome sequence (Lander et al. 2001; Venter et al. 2001), however, has attracted a great deal of interest from those devoted to the study of the evolution of our genes and genome. Previously, only genomes of non-vertebrate organisms had been sequenced. Although these genomes are good models for many aspects of gene and genome evolution, they offer only a limited insight into evolution within the vertebrate lineage. Our genome is several times larger than the genomes of previously sequenced organisms. Yet, it harbours only about double the number of genes seen in the fly. This curiosity is owing to the fact that most of the genome (some 95%) does not code for any genes. The reasons for the accumulation and maintenance of such quantities of non-coding DNA are yet to be elucidated. Coding protein genes also have special distributions. Just a glance into the pool of human genes reveals extreme patterns. While some genes are intronless, others contain

over 100000 bp of intronic sequence. In a narrower range is the variability found in coding sequence lengths. Most ribosomal proteins are less than 50 amino acids (aa) long, while the largest proteins are over 5000 aa long. Other sequence parameters such as base composition, codon usage, and amino acid distribution also show a great degree of variability. Gene location along chromosomes is also peculiar: while genes are tightly packed in some regions, large sections of the chromosomes are mostly devoid of genes.

By what processes have these features arisen? Are these patterns the result of neutral processes? Or are they partly accounted for by selective pressures related to gene activity? For example, is there any functional relationship between the number of genes and genome size? What is the role of transposable elements during genome evolution? And what accounts for the high degree of variability in gene characters? How does genome structure relate to gene location and gene regulation?

During my PhD I have devoted my time to examining the possible relationships between expression patterns with gene characteristics such as gene size, intron content, base composition, codon distributions, amino acid usage, and gene location along the chromosomes.

The evolution of genes is constrained to some extent by the requirements for the function of the proteins encoded. Because all protein coding genes are transcribed and translated we can hypothesise that reduction of costs related to protein synthesis would be favoured by selection. Were this the case then all genes would be under a general source of selective pressure. Because some genes are more frequently transcribed than others, we would expect the strength of selection to vary from gene to gene. Those genes that are frequently transcribed would therefore be under higher pressure to minimise protein

synthesis costs. Therefore, a relationship between expression levels and gene characteristics would be expected.

Evidence for such relationships between expression and gene characters has been recovered in unicellular species and invertebrates (for review see Akashi 2001). In those species, expression profiles of genes are related to gene features such as codon bias (Duret and Mouchiroud 1999; Gouy and Gautier 1982; Sharp et al. 1986; Stenico et al. 1994), intron (Vinogradov 2001b; Vinogradov 2001c) and coding (Moriyama and Powell 1998) region length, and amino acid usage (Akashi and Gojobori 2002). Compared to these species, mammalian populations have much smaller population sizes. Classical theory predicts that under such circumstances drift, rather than selection, is likely to be the dominant force in molecular evolution (for review see Kimura 1991). Therefore the observed distributions of mammalian gene characteristics are generally thought to be the result of selective pressure to preserve protein function and mutation processes but not affected by general selective pressures associated for example with protein efficiency. But do mammalian genes indeed not show any signs of expression mediated selection?

In **Chapter two** I analyse codon usage bias of human genes. Bias in usage of alternative codons has been observed in many species of bacteria (*Escherichia coli*, *Bacillus subtilis*) and unicellular (*Saccharomyces cerevisiae*) and invertebrate eukaryotes (*Caenorhabditis elegans*, *Drosophila melanogaster*). In these species codon bias has been found to be related to expression levels. Genes with higher expression tend to present higher bias towards the use of a particular set of codons (Duret and Mouchiroud 1999; Gouy and Gautier 1982; Sharp et al. 1986; Stenico et al. 1994). Although evidence that codon bias is the result of optimisation of protein synthesis is fairly well accepted, the exact nature of the optimised aspect is not a closed question (Powell and Moriyama 1997).

Codon bias has been mainly associated with selective pressures at the translational level. Codon usage may respond to 1. tRNA availabilities (Kanaya et al. 1999; Moriyama and Powell 1997; Sharp et al. 1995b), 2. accuracy in anticodon recognition (Grosjean and Fiers 1982) and 3. secondary structure of mRNA (Hartl et al. 1994). Codon usage bias could might also reflect selection for transcription efficiency. Vinogradov (Vinogradov 2001a; Vinogradov 2003), proposed that an increased G+C content in open reading frames would result in a more suitable DNA structure for transcription. This has yet to be proven but it is worth noticing that most of the preferred codons in Drosophila are G+C ending.

In addition, it should be noted that alternative explanations to selective pressures acting over synonymous sites, such as mutational processes and biased gene conversion have been put forward (Duret 2002). Neutral mutation processes associated with recombination appears to be a determinant factor in silent site base composition in drosophila and the nematode, thereby shaping codon usage patterns (Marais et al. 2001; Marais et al. 2003). Therefore the extent to which codon usage is related to translational selective pressures is still a matter of debate.

In mammals non-equal use of alternative codons has also been observed (Eyre-Walker 1991a; Wada et al. 1992). However, in mammalian genes codon bias is thought not to be influenced by translation-related selective pressures (Duret and Mouchiroud 2000). As noted above, because mammalian population sizes are relatively small, it is thought that variables related to protein translation puts very weak or no selective pressure on synonymous sites. Mammalian genomes have a great degree of heterogeneity in their base composition (isochore structures Bernardi 1993). This skew in nucleotide composition in regions of the genome affect gene sequences as well: the G+C concentration of coding regions correlate with G+C content of intronic (Urrutia and Hurst 2001) and intergenic

regions (Clay et al. 1996; Duret and Hurst 2001). Therefore the accepted view is that codon distributions in mammalian genes respond mainly to background or regional nucleotide composition (Bernardi 1995; Bernardi et al. 1997; Duret and Mouchiroud 2000).

In addition to the variations in nucleotide content across the genome, the choice of nucleotides in synonymous sites is affected by the identity of the adjacent bases. In several species, it has been observed that the frequencies of the 16 possible pairs of nucleotides are not equally represented in the genome -even when correcting for nucleotide content (Karlin and Mrazek 1996). This effect extends not only to pairs of synonymous sites with the adjacent bases, but also to the first two nucleotides of codons and to intron and intergenic regions.

All the above discussed effects make it difficult to assess the input of selection on codon bias in mammalian genes, but nonetheless the majority view is that selection is not an important factor for codon bias (Duret and Mouchiroud 2000; Eyre-Walker 1991b; Sharp et al. 1995a). There exist, however, a few studies that support a different position. Debry and Marzluff (1994) analysed rodent histone genes (which, note, are very highly expressed) and found evidence for a bias towards G+C ending codons over expectations from surrounding noncoding sequence. Similar conclusions were obtained by Iida and Akashi (2000) who examined codon distributions from constitutive and alternatively transcribed exons. Constitutive exons on average tended to have a larger proportion of G+C ending codons than those which may be spliced out in some isoforms. Further evidence for the importance of codon distributions for protein synthesis rates stems from molecular experiments where expression of foreign transfected genes into mammalian cells is increased by the substitution of rare codons by common ones in the mammalian genome (Levy et al. 1996; Zhou et al. 1999; Zolotukhin et al. 1996). These observations support the

case of selection-driven codon bias in mammalian genes. However, all of these studies were performed on small samples of genes or even single genes. In addition, genes were not randomly chosen; histones for example, are a very peculiar set of highly expressed genes, and therefore cannot be held as representative of the average gene. The results from experimental mammalisation of genes may constitute only the small proportion of cases (out of a lot of failures) where this procedure actually worked. The increased expression could be explained by changes in mRNA secondary structure or other gene-specific characters.

In Chapter two, with a sample of over 2000 human genes using a randomization protocol, I examine codon usage patterns. I evaluate whether human genes have a higher degree of codon bias than expected by their nucleotide distribution. I also examine the relationship between codon bias and breadth of expression and the input of dinucleotide biases. I show that human genes have a higher degree of codon bias from that expected by their base composition or dinucleotide distributions. In order to quantify the residual bias left after correcting for background nucleotide distributions, I present a new index. Many indexes to measure codon bias have been proposed (Karlin and Mrazek 1996; RodriguezBelmonte et al. 1996; Sharp and Li 1987; Wang et al. 1998; Wright 1990). However the majority of them require a known set of preferred codons. Of those not requiring predetermined major codons, ENc is the most widely used (Wright 1990). However this method assumes equiprobability of alternative codons as the null distribution and it is not suitable to use when nucleotides are not equally represented. Karlin and Mrazek (1996) proposed an alternative method, which allows one to test alternative expected distributions to that of equiprobability, but it is strongly dependent on amino acid distributions.

Given the limitations of the methods available for measuring codon usage bias, I therefore wished to develop an alternative method, that is both easy to calculate and that would allow correction for background nucleotide biases. The new method measures the degree of non-randomness in the use of alternative codons, and is minimally sensitive to differences in amino acid proportions of different degrees of degeneracy or rare amino acids. I denominated the method as Maximum likelihood Codon Bias (MCB) where the contribution to the index of the bias for each amino acid is weighed by an estimation of the likelihood of occurrence of bias on each amino acid, given its frequency and degree of degeneracy. Nevertheless, MCB is not a maximum likelihood method in a strict sense. I believe this method would be useful for interspecies comparisons by allowing correction for differences in nucleotide composition.

Using this method on the sample of 2000+ human genes, I performed an analysis of the relationship between codon bias and expression levels. However, the expression data available at the time was derived from EST data and only allowed me to calculate breadth of expression (number of tissues in which a gene is expressed), therefore limiting the interpretation of the results.

If indeed significant selective pressures do exist for optimising gene translation and or transcription, then we may expect expression patterns to be related not only to codon bias but to other gene characters as well. Until recently, comparable gene expression data for large numbers of genes was unavailable. With the release of the genome sequence, large scale expression profiling became possible. Currently, there are at least four sources of large scale expression data for human genes for several tissues.

1) **EST libraries** are derived from the sequencing of small cDNA fragments which are then paired with the corresponding gene. EST libraries were the only expression data available for a long time and are still the only record of expression profiles for many species and tissues. The main problem with this type of data is that some of the cDNA libraries obtained so far have been processed to eliminate redundant sequences and therefore quantification of expression levels is of limited validity. This process is called "normalization" and is performed by the hybridization of two duplicate libraries. More abundant would tend to hybridise more easily than rare transcripts. All annealed transcripts are then discarded and a random sample of cDNAs is selected for sequencing. In addition, because this method usually implies the sequencing of relatively long fragments of DNA, individual libraries are composed of small numbers of sequences which reduces the sample size and the reliability of the expression index estimates.

2) In **Serial Analysis of Gene Expression technology (SAGE)** (Lash et al. 2000; Velculescu et al. 1995) mRNAs bind to a column through their poly AA tails. Sequences are then cleaved with a restriction enzyme so that only the 3' end after the last restriction site of each sequence remains bound to the column and the rest of the fragments are washed away. Then the enzyme NIAIII is added to cleave at CATG sites, and fragments are washed away. As sequences are bound to the column by their poly AA tails, only the segment closer to the 3' end after the last restriction site remains attached to the column. Tags ten bases long are then cut and joined together with intermediate sequences so that direction is preserved. After sequencing, tags are paired to their putative gene based on identity with the gene coding sequence. SAGE technology allows the measurement of expression profiles for large numbers of genes in a relatively unbiased way by avoiding gene-specific

mRNA screening. For this reason, data from different laboratories can continue to be added. Over 150 human SAGE libraries are now available for more than 20 tissues.

SAGE data present two problems. First, in contrast to chip-array technology, although the method does not screen for particular genes, the fact that the recovered fragments are only ten bases long poses a problem for recognizing the corresponding gene. In principle, because ten bases can encode over 10 million different sequences there should be enough room to discriminate among genes. However, in practise many genes are recent duplicates of each other or share conserved functional domains and therefore only ten bases are not enough to distinguish them. In addition, a recent report by Margulies et al. (2001) showed that in some cases high G+C content sequences might be overrepresented since they are less likely to degrade in the process. The first obstacle is not of particular importance when only general statistics of a large number of genes are to be recovered, but might hamper the assessment of the expression patterns of certain individual genes. As for the G+C rich sequence bias, this can be corrected by identifying and excluding biased libraries. This may be done by the visual or statistical inspection of the skewness of the frequency distribution of sequences with respect to their G+C content (Margulies et al. 2001).

3) **Expression atlas** provides data for 12000 genes for over 45 human tissues and cell lines (Su et al. 2002). A high-density oligo-nucleotide array method using Affymetrix technology was used to compile this data set, which uses a set of 25 bp long oligo-nucleotides (one for each gene) printed over a glass slide. Purified mRNA is then passed over and binding quantities for each of the printed oligos is quantified. Because Affix printed slides are only read by Affix special readers, future additions to the dataset would require the use of the same equipment to facilitate correspondence between individual data

points. In addition, this method tests only for a particular set of genes and therefore, future additions to the dataset would require the use of the same set of primers. This method poses some problems as well. First, because oligo-nucleotides are not always specific enough to their target transcript some mRNAs may bind more than one print or none. Second, this method is likely to overestimate the number of genes expressed in each particular tissue since a minimum signal is almost always recovered for each gene. Therefore, a decision has to be made to set a minimum signal below which genes would be considered as non-expressed. The authors (Su et al. 2002) suggest setting a minimum value of 20 for a gene to be considered to be expressed. However, this threshold results in most genes being expressed in most tissues, in other words, most genes are constitutively expressed throughout the body. While this could be the case, this pattern is not consistent with the rest of available databases on gene expression so far produced. Therefore, it is probable that this particular database has a great chance of overestimating expression breadth. Expression rate estimates are largely unaffected by this. More recently, it has been reported that oligo printing order might influence expression levels recovered using chip technology (Balazsi et al. 2003). This artefact is potentially serious if a particular order of genes is used in the printing process (i.e. chromosomal position).

4) Bodymap data was collected by the sequencing of 3' ends of purified mRNAs (Hishiki et al. 2000; Kawamoto et al. 2000). This method may provide higher assurance of gene identity as it involves sequencing of a longer sequence from the mRNA than that used in the SAGE method. However, because of the fact that it is time consuming, only a limited number of sequences were obtained. This makes difficult the assessment of expression patterns for large numbers of genes.

Bodymap was not analysed in the manuscripts that make up this thesis because of the small sample size of their libraries. EST data are suitable for estimating breadth of expression (number of tissues where a gene is expressed): however, as most of the libraries are normalised, quantification of levels of expression is not easily obtained and may not be reliable. SAGE and chip array data allow quantification of expression. From these two datasets three indexes of expression levels for each gene can be obtained: a) peak expression –the highest expression of a gene in any given tissue-, b) mean or level of expression –average expression of a gene in all tissues where it is expressed- and c) breadth of expression –the number of tissues where a gene is expressed. Because all of the datasets are subject to some degree of noise and biases in expression quantification, more than one dataset was used for the analyses presented. In this way, the chances of artefactual relations of expression and gene characteristics were reduced. It should be noted that, in all cases, similar results were obtained when different datasets were examined.

Chapter three and onwards make use of these expression datasets to analyse the relationship between expression patterns of genes and their characteristics. **Chapter three,** in particular, analyses the influence of expression levels on gene sequence characteristics such as protein length, intron content, codon bias and amino acid composition. If highly expressed genes optimise protein synthesis costs, then we expect them to have a greater degree of codon bias, to minimise the length of transcribed and translated sequences and to encode for metabolically cheaper amino acids. In the case of codon bias, there is an extensive line of evidence pointing towards a relationship with expression levels in non-vertebrate species.

In the case of gene length there is more scarce or indirect evidence for a relation with expression levels. In yeast and Drosophila, protein length is negatively related to

codon bias (Moriyama and Powell 1998). A similar result was also found when comparing intron length with codon bias (Vinogradov 2001b; Vinogradov 2001c). Assuming that codon bias is an indirect estimate of expression levels, these observations suggest that gene length is under pressure to be reduced (or not increased) in highly expressed genes.

I analysed the nature of the relationship between expression patterns and gene length of both coding and non coding regions. To correct for possible regional effects, in my analysis I calculated intergenic distances for all annotated genes in contigs and calculated G+C content from 5000 bp long intergenic fragments. These data were then used to assess whether intron and coding region sizes are dependent on these regional genome characteristics. Results from this analysis show that there is indeed a strong dependence of intron length on intergenic distance and the base composition of surrounding sequence. Multiple regression analyses were used to evaluate the influence of expression levels over gene characteristics such as intron and coding regions length, codon usage bias and amino acid composition, after taking into account regional effects.

The observation of the correlation of gene sequence parameters with the characteristics of the chromosome region where they are located, suggested a possible relationship between expression patterns and genomic location. In **Chapter four** I analyse in some detail the sorting of genes with respect to their expression patterns. The analysis, based on SAGE data for over 10000 genes, found that broadly expressed genes were found to be situated among other broadly expressed genes. However the reasons for this clustering remained unknown. One possibility is that these clusters respond to some sort of co-regulation. In the human genome, there is little evidence that operon-like structures are the norm. However, identification of regulatory elements is still precarious for most genes.

Alternatively they could respond to structural genomic properties. **Chapter 5** presents detailed analysis of gene expression and gene distribution with respect to genome structure.

Together the analyses presented in this thesis address different aspects in which selective pressures related to protein synthesis cost-optimisation might influence gene sequence characteristics, in particular of highly expressed genes.

Bibliography

Akashi, H. 2001. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11:** 660-666.

Akashi, H. and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. USA* **99:** 3695-3700.

Balazsi, G., K.A. Kay, A.L. Barabasi, and Z.N. Oltvai. 2003. Spurious spatial periodicity of co-expression in microarray data due to printing design. *Nucleic Acids Res* **31:** 4425-4433.

Bernardi, G. 1993. The Isochore Organization of the Human Genome and Its Evolutionary History - a Review. *Gene* **135:** 57-66.

---. 1995. The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* **29:** 445-476.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1997. Isochores and synonymous substitutions in mammalian genes. In *DNA and Protein Sequence Analysis* (eds. M.J. Bishop and C.J. Rawlings). IRL Press, Oxford.

Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud, and G. Bernardi. 1996. Human coding and noncoding DNA: Compositional correlations. *Mol. Phylogenet. Evol.* **5:** 2-12.

Debry, R.W. and W.F. Marzluff. 1994. Selection on Silent Sites in the Rodent H3 Histone Gene Family. *Genetics* **138:** 191-202.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12:** 640-649.

Duret, L. and L.D. Hurst. 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* **18:** 757-762.

Duret, L. and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. *Proc. Natl Acad. Sci. U.S.A.* **96:** 4482-4487.

---. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17:** 68-74.

Eyre-Walker, A. 1991a. An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* **33:** 442-449.

Eyre-Walker, A.C. 1991b. An analysis of codon usage in mammals: selection or mutation bias? *J Mol Evol* **33:** 442-449.

Gouy, M. and C. Gautier. 1982. Codon usage in bacteria - correlation with gene expressivity. *Nucleic Acids Res* **10:** 7055-7074.

Grosjean, H. and W. Fiers. 1982. Preferential codon usage in prokaryotic genes - the optimal codon anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18:** 199-209.

Hartl, D.L., E.N. Moriyama, and S.A. Sawyer. 1994. Selection Intensity for Codon Bias. *Genetics* **138:** 227-234.

Hishiki, T., S. Kawamoto, S. Morishita, and K. Okubo. 2000. BodyMap: a human and mouse gene expression database. *Nucleic Acids Res* **28:** 136-138.

Iida, K. and H. Akashi. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261:** 93-105.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238:** 143-155.

Karlin, S. and J. Mrazek. 1996. What drives codon choices in human genes? *J. Mol. Biol.* **262:** 459-472.

Kawamoto, S., J. Yoshii, K. Mizuno, K. Ito, Y. Miyamoto, T. Ohnishi, R. Matoba, N. Hori, Y. Matsumoto, T. Okumura et al. 2000. BodyMap: a collection of 3' ESTs for analysis of human gene expression information. *Genome Res* **10:** 1817-1827.

Kimura, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet* **66:** 367-386.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lash, A.E., C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, and S.F. Altschul. 2000. SAGEmap: A public gene expression resource. *Genome Research* **10:** 1051-1060.

Levy, J.P., R.R. Muldoon, S. Zolotukhin, and C.J. Link. 1996. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat. Biotechnol.* **14:** 610-614.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc. Natl Acad. Sci. U.S.A.* **98:** 5688-5692.

---. 2003. Neutral effect of recombination on base composition in Drosophila. *Genet Res* **81:** 79-87.

Margulies, E., S. Kardia, and J. Innis. 2001. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res* **29:** e60.

Moriyama, E.N. and J.R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. *J. Mol. Evol.* **45:** 514-523.

---. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Res* **26:** 3188-3193.

Powell, J.R. and E.N. Moriyama. 1997. Evolution of codon usage bias in Drosophila. *Proc. Natl Acad. Sci. U.S.A.* **94:** 7784-7790.

RodriguezBelmonte, E., M.A. FreirePicos, A.M. RodriguezTorres, M.I. GonzalezSiso, M.E. Cerdan, and L.M. RodriguezSeijo. 1996. PICDI, a simple program for codon bias calculation. *Mol. Biotechnol.* **5:** 191-195.

Sharp, P.M., M. Averof, A.T. Lloyd, G. Matassi, and J.F. Peden. 1995a. DNA sequence evolution: the sounds of silence. *Philos Trans R Soc Lond B Biol Sci* **349:** 241-247.

---. 1995b. DNA-Sequence Evolution - the Sounds of Silence. *Philos. Trans. R. Soc. Lond. B* **349:** 241-247.

Sharp, P.M. and W.H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15:** 1281-1295.

Sharp, P.M., T.M.F. Tuohy, and K.R. Mosurski. 1986. Codon usage in yeast - cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14:** 5125-5143.

Stenico, M., A.T. Lloyd, and P.M. Sharp. 1994. Codon usage in caenorhabditis-elegans - delineation of translational selection and mutational biases. *Nucleic Acids Res* **22:** 2437-2446.

Su, A.I., M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. U.S.A.* **99:** 4465-4470.

Urrutia, A.O. and L.D. Hurst. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159:** 1191-1199.

Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler. 1995. Serial Analysis of Gene-Expression. *Science* **270:** 484-487.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Vinogradov, A.E. 2001a. Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* **18:** 2195-2200.

---. 2001b. Intron length and codon usage. *J. Mol. Evol.* **52:** 2-5.

---. 2001c. Intron length and codon usage (vol 52, pg 2, 2001). *J. Mol. Evol.* **52:** 310-310.

---. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res* **31:** 1838-1844.

Wada, K.N., Y. Wada, F. Ishibashi, T. Gojobori, and T. Ikemura. 1992. Codon Usage Tabulated from the Genbank Genetic Sequence Data. *Nucleic Acids Res* **20:** 2111-2118.

Wang, T.T., W.C. Cheng, and B.H. Lee. 1998. A simple program to calculate codon bias index. *Mol. Biotechnol.* **10:** 103-106.

Wright, F. 1990. The Effective Number of Codons Used in a Gene. *Gene* **87:** 23-29.

Zhou, J., W.J. Liu, S.W. Peng, X.Y. Sun, and I. Frazer. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J. Virol.* **73:** 4972-4982.

Zolotukhin, S., M. Potter, W.W. Hauswirth, J. Guy, and N. Muzyczka. 1996. A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* **70:** 4646-4654.

Chapter two

Codon Usage Bias in Human Genes

Urrutia, A. O., and L. D. Hurst. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159: 1191-1199.

Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection

Araxi O. Urrutia and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, United Kingdom

Manuscript received May 22, 2001
Accepted for publication August 15, 2001

ABSTRACT

In numerous species, from bacteria to Drosophila, evidence suggests that selection acts even on synonymous codon usage: codon bias is greater in more abundantly expressed genes, the rate of synonymous evolution is lower in genes with greater codon bias, and there is consistency between genes in the same species in which codons are preferred. In contrast, in mammals, while nonequal use of alternative codons is observed, the bias is attributed to the background variance in nucleotide concentrations, reflected in the similar nucleotide composition of flanking noncoding and exonic third sites. However, a systematic examination of the covariants of codon usage controlling for background nucleotide content has yet to be performed. Here we present a new method to measure codon bias that corrects for background nucleotide content and apply this to 2396 human genes. Nearly all (99%) exhibit a higher amount of codon bias than expected by chance. The patterns associated with selectively driven codon bias are weakly recovered: Broadly expressed genes have a higher level of bias than do tissue-specific genes, the bias is higher for genes with lower rates of synonymous substitutions, and certain codons are repeatedly preferred. However, while these patterns are suggestive, the first two patterns appear to be methodological artifacts. The last pattern reflects in part biases in usage of nucleotide pairs. We conclude that we find no evidence for selection on codon usage in humans.

DOES selection act on mutations within exons that do not alter the amino acid sequence of the coded protein? Originally it was asserted that these synonymous mutations must be neutral (KING and JUKES 1969). However, it is well known that unequal use of alternative codons is a common phenomenon in many unicellular species, as well as in Drosophila and Caenorhabditis, and that this may reflect the activity of selection (MARAIS *et al.* 2001). In several unicellular species (GOUY and GAUTIER 1982; SHARP *et al.* 1986; STENICO *et al.* 1994) and some invertebrates (DURET and MOUCHIROUD 1999) it has been observed that codon usage bias is related to expression patterns. In these species it has been found that the extent of codon usage bias correlates with levels of gene expression, where highly expressed genes tend to have a greater bias (GOUY and GAUTIER 1982; SHARP *et al.* 1986; STENICO *et al.* 1994; DURET and MOUCHIROUD 1999). Evidence has been found to relate this to tRNA availabilities (SHARP *et al.* 1995; MORIYAMA and POWELL 1997; KANAYA *et al.* 1999). These observations suggest that in these species codon usage bias is explained partly by translation efficiency-related pressures. Additionally, the rate of synonymous evolution covaries with the level of codon bias (see, *e.g.*, POWELL and MORIYAMA 1997), although this might be an artifact of the method (DUNN *et al.* 2001). There are also consistent preferences toward certain codons within any given genome that may be interpreted as a result of selection as they appear to be in the opposite direction to mutation bias.

In mammalian genomes, codon usage bias is also observed (EYRE-WALKER 1991; IKEMURA and WADA 1991). However, as mammalian genomes show a great variation in nucleotide concentrations across the genome (*i.e.*, isochores; BERNARDI 1995; BERNARDI *et al.* 1997), codon usage bias has been attributed to this background nucleotide bias. In Figure 1, for example, we plot GC content of third sites in exons for 369 genes on human chromosomes 21 and 22, against the GC content of the 50 kb of DNA flanking the gene in question. As can be seen the two strongly covary. A similar covariance is also seen between GC content at third sites and intronic GC (see DURET and HURST 2001 and references therein). As codon usage bias strongly reflects the GC content at the silent sites (Figure 2a), it has been difficult to assess the input of other variables to bias in codon usage in mammalian genes related to protein synthesis such as expression patterns and rates of substitutions.

Nonetheless in one case there has been a claim that a highly expressed set of genes (histones) does show codon usage that deviates from background nucleotide content in flanking regions (DEBRY and MARZLUFF 1994). This suggests that selection could operate on codon usage in humans as well. If this were generally true we might expect that codon usage bias in mammals should influence expression patterns. It is unknown

Corresponding author: Laurence D. Hurst, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom. E-mail: l.d.hurst@bath.ac.uk

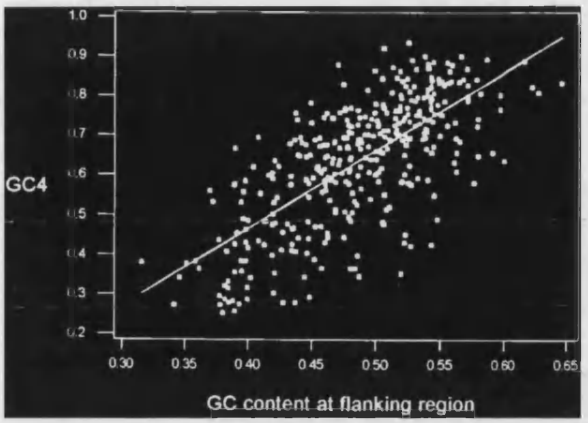


FIGURE 1.—Correlation of GC content at fourfold degenerate sites with the GC content at 50-kb flanking region. (GC4 = −0.32 + 1.96 GC50 kb; $r^2 = 50.4\%$; $P < 0.001$).

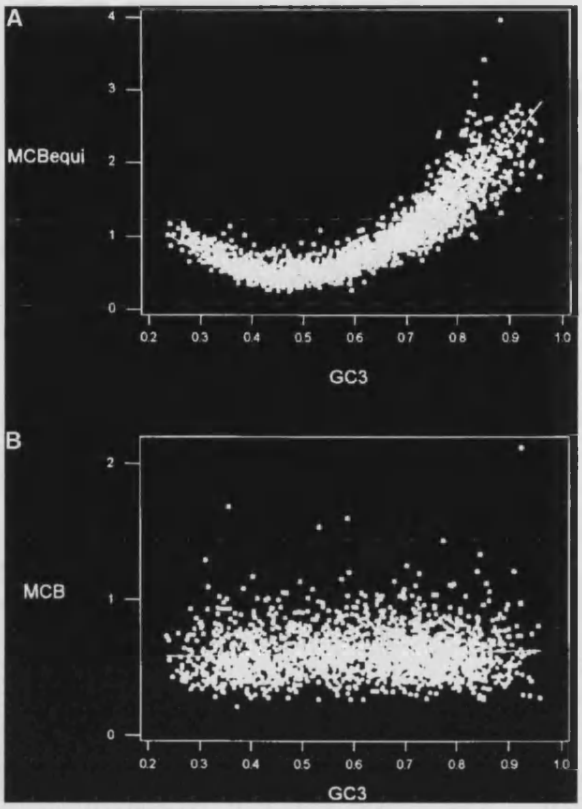


FIGURE 2.—Correlation of codon usage bias using the MCB method and G + C content at third sites (GC3s), (A) assuming equiprobability and (B) correcting by nucleotide bias in a sample of 2396 human genes. (MCB = 0.57 + 0.067 GC3s, $r^2 = 0.5\%$, $P = 0.001$; MCBequiprobability = 2.61 − 8.94 GC3s + 9.56 × $GC3s^2$, $r^2 = 85.3\%$, $P < 0.001$).

whether, more generally, codon usage bias, after correction for background nucleotide content, covaries with any expression parameters. By contrast, there is now some evidence supporting the notion that codon usage affects expression. When nonmammalian genes are to be expressed in mammalian cells, the replacing of rare codons in the mammalian genome for common ones appears to have dramatic effects on the level of gene expression. This method, known as "mammalianization" or "humanization," has been used for increasing the expression of several genes (*e.g.*, LEVY *et al.* 1996; ZOLOTUKHIN *et al.* 1996; WELLS *et al.* 1999; ZHOU *et al.* 1999).

In this study we present the results from the analysis of codon usage bias in a sample of over 2000 human genes designed to ask whether codon usage bias in mammals can be explained by background nucleotide content alone or whether such parameters as expression breadth might also be important. To achieve this we developed a tool to measure codon usage bias, correcting nucleotide biases.

METHODS

Sequences from 2396 genes were included in the sample. Accession numbers were obtained from the DURET and MOURICHOUD (2000) database and sequences retrieved by ACNUC (GOUY *et al.* 1984). All incomplete sequences (*i.e.*, with internal gaps, nondefined nucleotides, or no start or stop codon) were discarded. Data of expression patterns were also obtained from the DURET and MOUCHIROUD (2000) database. Breadth of expression was calculated by counting the number of tissues where the gene is expressed. Columns referring to the same tissue but in different developmental stages were treated as a single tissue.

Randomization tests: Random sequences were generated for each gene conserving the base content at first, second, and third sites and for gene length. Start and stop codons were removed from the randomizations. During randomizations all sequences that contained an internal stop codon were discarded. The procedure was repeated until a total of 1000 random sequences were obtained.

Effective number of codons tests: Effective number of codons (ENC) values were obtained for all sequences and values of original sequences were compared with the distribution of random sequences. As the ENC index has a cutoff at 61 and all sequences with greater values are adjusted to 61, the variance of the distribution was estimated on the basis of the median instead of the mean and by using only the lower half of the distribution.

Defining amino acids: In all tests, nondegenerative amino acids (methionine and tryptophan) were not taken into account. For the majority of the amino acids all their alternative codons have the same bases at the first and second site. The exceptions are serine, arginine, and leucine, each encoded by six alternative codons.

Each of these amino acids was treated in all tests as two independent amino acids, one of twofold degeneracy and one of fourfold degeneracy.

Background nucleotide bias model expectations: To obtain expected proportions for each alternative codon correcting for background nucleotide content, all codons were split into three groups according to the number of different nucleotides (two, three, and four) that could appear at the third site without changing the amino acid encoded. The group of degeneracy two was further divided into two groups, those where the choice is between T and C and those ending in A or G. The expected proportions of each alternative codon for a given amino acid were derived from all the other sites with the same degree of degeneracy or greater (*i.e.*, excluding the amino acid being analyzed). For example, for amino acids with two degrees of degeneracy that could use the nucleotides thiamine and cytosine at the third site, expectations were calculated on the basis of all the other amino acids of two degrees of degeneracy that had a choice of the same nucleotides for the third site and also all the amino acids of four degrees of degeneracy were included by calculating the relative frequencies of thiamine and cytosine. For isoleucine, expectations were calculated by calculating the relative frequencies of adenine, cytosine, and thiamine in fourfold degenerate amino acids. Finally, for all the fourfold degenerate amino acids, only the distributions of nucleotides at the third sites of other fourfold degenerate sites were used for calculating expectations.

To minimize the uncertainty in the expected values, all cases with <30 sites to base the expectations on were eliminated. It should be noted that by using this model as null expectation, we are not taking into account the codon bias caused by dinucleotide biases.

Probability of observed bias: Proportions of observed and expected codon usage for each amino acid were represented in terms of the minimal number of binomial variables. For amino acids with two alternative codons, codon usage is represented in terms of one variable, the frequency of one codon over the number of times the amino acid is present. For three alternative codons A, B, and C, codon usage can be represented with two variables: (a) the proportion of codon A over the total number of times the amino acid is present and (b) the proportion of codon B from the sum of frequencies of codons B and C. For amino acids with four alternative codons A, B, C, and D, the proportions of codon usage are represented by three variables: (a) the proportion of codons A + B over the frequency of the amino acid, (b) the proportion of codon A over the sum of frequencies of codons A + C, and (c) the proportion of B over the sum of frequencies of codons B + D. Under this method, the distribution of codon usage of a gene, and the expected one, can be represented by 38 binomial variables. All sequences in which not all the variables could be assessed were excluded from analysis (leaving $n = 1629$). To estimate the probability of the bias observed for each gene under the null hypothesis, the deviation from expectation for each variable was represented in terms of the numbers of standard deviations away from the mean (z). The standard deviation for a binomial variable can be defined as

$$\sigma = \sqrt{\frac{P(p) \cdot P(q)}{N}}.$$

The squared z values for each of the 38 variables were calculated and then summed to obtain the overall score (x),

$$z = \frac{O - E}{\sigma}$$

$$x = \sum z^2.$$

Assuming that the binomial variables are normally distributed, the probability of occurrence of the observed bias can then be calculated with a χ^2 distribution of 38 d.f. that has the following probability density function:

$$f_x(x; 38) = \frac{1}{2^{38/2} \cdot \Gamma(38/2)} x^{(38/2)} \cdot e^{-x/2}.$$

Analysis of overall bias: All sequences were concatenated into a single large sequence. Observed and expected codon distributions were obtained as previously described for individual genes. The probability of the observed bias or greater from background nucleotide bias model expectations for each amino acid with n alternative codons was estimated using two standard methods of goodness of fit that approximate the χ^2 distribution with $n - 1$ d.f.:

A. χ^2 test:

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

B. *G*-test:

$$G = 2 \cdot \sum (O \cdot \ln(O/E)).$$

Probabilities were estimated for the values obtained by comparing with cumulative χ^2 distributions.

Dinucleotide analysis: Dinucleotide proportions were obtained for sites first to second, second to third, and third to first for each gene. Expected proportions for a given dinucleotide (E_d) at sites s2 and s3 given the nucleotide content in the sequence at each site were calculated as

$$E_{d,s2,s3} = p(n_2) \cdot p(n_3),$$

where $p(n_2)$ and $p(n_3)$ are the proportions of the second and third nucleotides of the dinucleotide at sites s2 and s3. Dinucleotide bias (DnB) was estimated for each gene by

$$\text{DnB} = \sum(E_{d,s2,s3} - O_{d,s2,s3})^2,$$

where $O_{d,s2,s3}$ is the observed proportion of the dinucleotide.

RESULTS

Background nucleotide content alone does not explain the codon bias observed in mammalian genes: The extent of codon usage bias in human genes is dominantly dictated by the nucleotide content of the chromosomal region within which the gene finds itself (BERNARDI 1995). Does this alone explain the degree of codon usage bias? We studied codon usage bias in a sample of 2396 human genes. As a first approach to investigate whether the observed codon bias can be explained by nucleotide biases at synonymous sites, 1000 random sequences were generated for each gene, conserving gene length and the base content per site. ENC values (WRIGHT 1990) were obtained as a measure of codon usage bias for original and random sequences. We then compared the ENC for the real sequences to the distribution of random sequences. Over one-half of the genes were more deviated than any of the random sequences generated and 81% were significantly deviated with $\alpha = 0.05$ (see METHODS).

However, because the ENC has a cutoff at 61 it has limited use for sequences with low codon usage bias. In addition, randomizing the first and second positions could potentially influence the distribution of ENC values in the random sequences. We therefore performed a second test to estimate the probability of the bias observed, by comparing the proportions of alternative codons of each sequence from the same sample of genes to expected proportions based on the nucleotide content at the third sites of each sequence. If the bias from equiprobability of a gene can be explained by the nucleotide content of that gene, then the null expectation would be that frequencies for each codon should match proportions of bases at the third site of all the amino acids with the same degree of degeneracy in that gene. To estimate the probability of obtaining the observed bias or greater under the null hypothesis for each gene, the observed and expected frequencies of codon usage for each amino acid were represented in terms of the minimal set of binomial variables, which is a method to approximate a multinomial distribution (see METHODS). The probability of obtaining the observed bias or greater under null expectation was estimated by summing the squared z values of distances of observed from expected for each binomial variable and comparing this with a standard χ^2 distribution (see METHODS). Significant deviations from expected (defined as $P < 0.01$) were found for 99% of the genes in the sample (data not shown). We conclude that "background" nucleotide content explains some, but by no means all, of the observed codon usage bias. Dinucleotide biases also are known to affect the bias in codon usage (HANAI and WADA 1988; KARLIN and MRAZEK 1996) so it was expected that some proportion of the genes would be more biased than expected by the background nucleotide content.

Prior methods for assessing codon usage bias have limitations: There are several methods to measure codon usage bias; however, many of them require a known set of preferred codons estimated from highly expressed genes. ENC (WRIGHT 1990) is a popular method that does not assume preferred codons but is not especially suitable for statistical analysis, as it does not allow testing null hypotheses for codon usage distribution other than equiprobability. KARLIN and MRAZEK (1996) proposed an alternative method that permits the introduction of values of an expected distribution. However, we found that this method is sensitive to biases on the use of amino acids of different degrees of degeneracy; *i.e.*, the proportion of fourfold degenerate amino acids of a sequence correlates with the index of codon usage bias ($r^2 = 14.1\%$; Figure 3a).

Maximum-likelihood codon bias is a new method for determining codon usage bias correcting for background nucleotide content: Given the limitations of the available methods, we chose to develop an alternative method that is easy to obtain and not sensitive to amino acid biases. We wanted a method that could measure the degree of nonrandomness in the use of alternative codons that is minimally affected by the presence of rare amino acids. In addition the method should allow testing of a variety of null hypotheses for codon distribution (*i.e.*, not just equiprobability of occurrence); in this article we use this method to correct for background nucleotide content, but it can be used to correct for dinucleotide biases as well.

The use of alternative codons can be thought of as an ensemble of several random variables, one per amino acid, each with two to six possible different outcomes or codons (amino acids encoded by only one codon cannot have codon usage bias), and each outcome with an associated probability of appearance. Each specific distribution of outcomes is a vector and the codon bias for one amino acid is the distance of the observed vector from the expected one. However, to obtain an index of codon usage bias for a complete gene, the biases of individual amino acids have to be added in a sensible way. Different amino acids within a gene vary in two aspects: frequency within a sequence and their degree of degeneracy. If an amino acid is rare, then the observed distribution is more likely to be far from the expected just by chance; therefore the bias of a rare amino acid should be downscaled to have less impact on the overall index of codon bias. The different amino acids also vary in the number of alternative codons by which they are encoded and this should also be taken into account when biases from different amino acids are to be added.

Taking into account the two aspects discussed above,

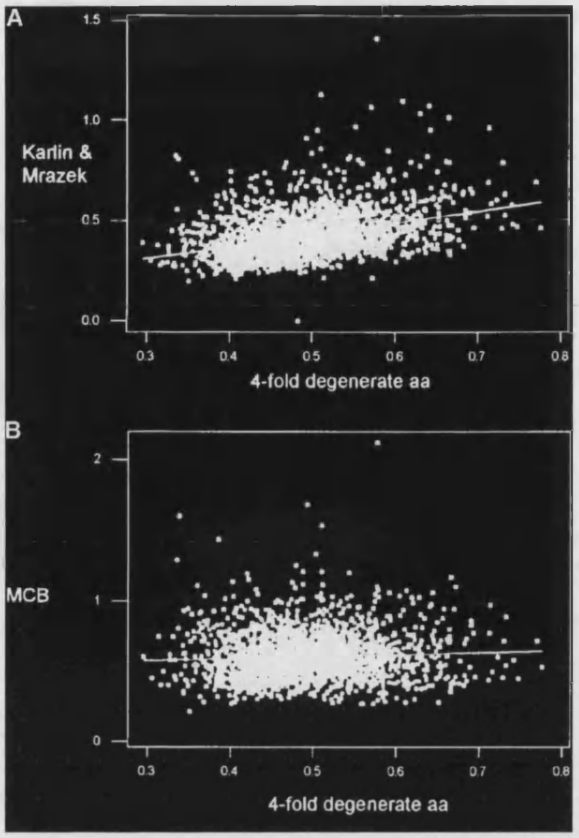


FIGURE 3.—Correlation of codon usage bias and the proportion of fourfold degenerate sites in the sequence, using (A) the Karlin and Mrazek (K&M) method (K&M = 0.136 + 0.588 fourfold degeneracy, $r^2 = 14.1\%$; $P < 0.001$) and (B) the MCB method (MCB = 0.535 + 0.145 fourfold degeneracy, $r^2 = 0.4\%$; $P = 0.002$).

we developed a new method that is easy to calculate and allows us to test different models to explain codon usage bias. The bias of an individual amino acid B_A with frequency N_A of level of degeneracy T, having the observed O_c and expected E_c proportions for each alternative codon, is obtained by

$$B_{AT} = \sum_c \frac{(O_c - E_c)^2}{E_c}.$$

The bias for a gene B_g can then be obtained by summing over all amino acids,

$$B_g = \sum \frac{B_{AT} \cdot \log N_A}{A},$$

where A is the number of amino acids contributing to the index.

All genes where more than five amino acids were missing or no index could be estimated were removed from all comparisons (leaving $n = 2387$). We denominated the method as maximum-likelihood codon bias (MCB), where the contribution to the index of the bias of each amino acid is weighted by an estimation of the likelihood of occurrence of bias on each amino acid, given its frequency and degree of degeneracy. Nevertheless, MCB is not a maximum-likelihood method in a strict sense. We believe this method would be useful for interspecies comparisons by allowing correction for differences in nucleotide composition. Importantly, MCB is minimally affected by biases in amino acid content of different degrees of degeneracy ($r^2 = 0.4\%$; Figure 3b) and appears to effectively remove the influence of background GC content (compare Figure 2A and 2B).

It should be noted that with any procedure that estimates the distance from randomness, the size of the sample of events affects the variance that is expected; since the length of genes varies it is expected that this would influence the MCB values that are obtained. Therefore it is important to carefully study the relation of gene length with the variables that are being tested against codon usage values. A script for calculating MCB is available from the authors.

MCB covaries with breadth of expression and rates of synonymous substitution: Expected distributions for each codon family were derived from the base composition of all third sites with the same or greater level of degeneracy within a given sequence (according to METHODS) and MCB values were obtained for all genes in the data sample. If the residual biases in codon usage, once correcting for nucleotide content, are due to selection then we could expect (a) higher bias in more broadly expressed genes, (b) consistently preferred codons, or (c) an inverse correlation with levels of synonymous substitutions (K_s).

We assessed the effect of breadth of expression on codon usage bias in our sample (see METHODS). Breadth of expression is not a direct measure of expression rate and therefore we may not necessarily be analyzing the key parameter. Nonetheless, the breadth of expression is known to covary with the intensity of purifying selection acting on the nonsynonymous sites (DURET and MOUCHIROUD 2000), so may reasonably be taken as a covariate to the strength of purifying selection. To assess the interaction between breadth of expression and codon usage bias the sample was divided into three groups according to the number of tissues in which they are expressed: (1) genes expressed in up to 5 tissues ($n = 1242$), (2) genes expressed in more than 5 but not more than 10 tissues ($n = 494$), and (3) genes expressed in between 11 and 15 tissues ($n = 272$). The levels of codon bias in the three groups were significantly different from each other (5 *vs.* 10, $P = 0.001$; 10 *vs.* 15, $P < 0.001$; 5 *vs.* 15, $P < 0.001$; Kruskal-Wallis test). In all comparisons genes expressed in fewer tissues tend to have a lower MCB value. The correlation line between MCB and the number of tissues is consistent with this result

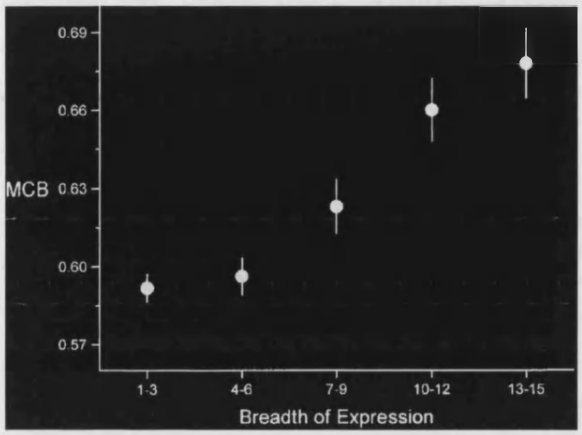


FIGURE 4.—The relationship between MCB and expression patterns. Average values of MCB and standard error bars are shown for the genes divided into five groups according to the number of tissues where they are expressed: 1–3, 4–6, 7–9, 10–12, and 13–15 (n = 884, 484, 293, 203, and 144, respectively).

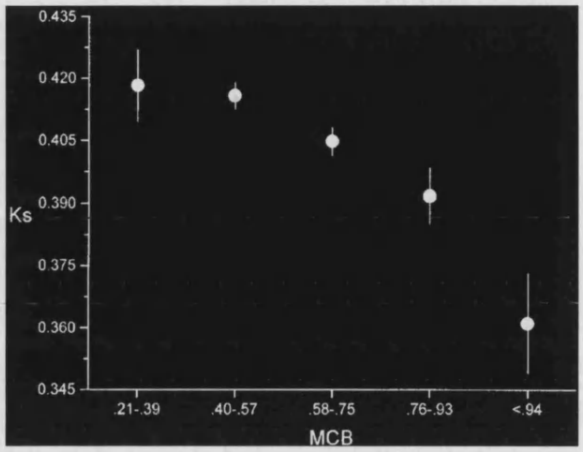


FIGURE 5.—The relationship between the rate of silent site evolution (K_s) and codon usage bias. Average value and standard error bars of K_s (human-rodent comparison using the Li method; DURET and MOUCHIROUD 2000) are shown for genes grouped by MCB values: 0.21–0.39, 0.40–0.57, 0.58–0.75, 0.76–0.93, and <0.94 (n = 153, 938, 874, 264, and 67, respectively).

($P < 0.001$, $r^2 = 3.1\%$; see Figure 4). This result suggests that genes with broader expression show a higher degree of codon usage bias. These results cannot be explained by compositional biases caused by transcriptional coupled mutational biases (*e.g.*, higher rate of C → T mutations in "breathing DNA") since the MCB method already takes into account gene-specific background nucleotide concentrations.

An inverse correlation between codon usage bias and rates of silent site substitutions has been observed in bacteria (SHARP and LI 1987), Drosophila (POWELL and MORIYAMA 1997), and yeast (L. D. HURST, unpublished data). If codon usage bias is due to selective pressures then it is expected that genes with higher codon usage bias would have lower rates of synonymous substitutions, although the effect may be weak. When rates of synonymous substitutions (compared to mouse and rat orthologs; DURET and MOUCHIROUD 2000), using LI's (1993) method (Li93) and removing tandem substitutions (data as in DURET and MOUCHIROUD 1999), were plotted against MCB values, we observed an inverse correlation between rates of silent site substitution and MCB values ($r^2 = 1.2\%$, $P < 0.001$; see Figure 5). A similar result is obtained ($r^2 = 1.4\%$, $P < 0.001$) when comparing MCB values with rates of substitution at the fourfold degenerate sites, using the Tamura and Nei protocol after removing tandem substitutions. While this result is consistent with selection, it must be treated with caution owing to the fact that estimators of K_s may be biased when nucleotide content is biased (DUNN *et al.* 2001). Indeed, the correlation is not present (or at most only weakly suggested) if instead we apply the maximum-likelihood method of Goldman and Yang ($P = 0.056$, $r^2 = 0.2\%$).

The covariance with expression breadth and synonymous substitution rates is also found when correlating the Karlin and Mrazek method as a measure of codon bias (breadth of expression, $P < 0.001$, $r^2 = 0.9\%$; synonymous substitution rate using Li93, $P = 0.002$, $r^2 = 0.4\%$). Although both the MCB and Karlin and Mrazek methods significantly correlate with synonymous rates of substitutions and breadth of expression, these weak correlations should be interpreted cautiously.

Codon preferences: The above results are suggestive of a role for selection. If selection is to explain the above effects then we should also expect to see certain codons repeatedly being favored among genes. To investigate if the observed biases were favoring specific codons over others we performed an overall analysis of the whole sample by concatenating all genes into one large sequence. If the biases are due to factors specific for individual genes these should cancel each other out in the whole sample. The proportions for each alternative codon were obtained and compared with expectations from the nucleotide biases. Significant differences from expectations were observed for all of the amino acids that have two or more synonymous codons using the two tests of goodness of fit ($P < 0.001$, see METHODS).

A more conservative test is to investigate the consistency of the direction of the biases for individual genes. If there is no significant tendency favoring a particular set of codons then it is expected that a codon would be overrepresented one-half of the times that it deviates from expectation. The majority of the codons have significantly less heterogeneity than expected by chance and some were biased in one direction in 90% of the

TABLE 1

Heterogeneity of direction of bias for each codon

First codon position	Second codon position U	C	A	G	Third codon position
U	(+)	++	(+)	(+)	U
C	−−	++	(−)	−−	
A	+	(+)	(+)	(−)	
G	−−	++	+	−−	
U	(−)	+	(−)	(−)	C
C	−−−	=	(+)	(−)	
A	+ +	(+)	(−)	(+)	
G	−−−	+ +	−	(+)	
U	−−	(+)	STOP	STOP	A
C	−−−	++	−−−	=	
A	−−−	++	(−)	+	
G	−−−	+	=	+	
U	++	−−−	STOP	TRP	G
C	+++	−−−	+++	+	
A	MET	−−−	(+)	−	
G	+++	−−−	=	−	

The degree of heterogeneity in the direction of bias toward under- or overrepresentation for each codon is shown. To facilitate interpretation, values were substituted by symbols. =, no significant deviation from expectation; + and −, significant over/under representation compared to expectation ($P < 0.01$); (+) and (−), significant deviation ($P < 0.01$) up to 10 standard deviations away from expected distribution; +/−, ++/−−, +++/−−−, between 10–20, 20–30, and <30 standard deviations.

genes (Table 1). The above results are consistent with selective pressures favoring specific codons. Were this the result of selection we can predict that tRNA levels should be more highly skewed for the amino acids showing bias than for those showing little bias as has been shown for other species (*cf.* SHARP *et al.* 1995; MORIYAMA and POWELL 1997; KANAYA *et al.* 1999); however, LANDER *et al.* (2001) did not find support for this prediction in human genes.

Expression breadth and synonymous substitution patterns are most probably due to gene length effects: The above results are suggestive of selection possibly playing a role in codon usage bias in humans. However, as stated earlier, genes of different length are likely to have different MCB values owing to the nature of the method. Indeed, if we randomize our sequences and measure the mean MCB for 1000 simulants for each of our genes, we find that the MCB, on average, is higher for shorter genes. This is to be expected of any statistic that employs a multinomial distribution and applies equally to the method of Karlin and Mrazek.

Importantly, it so happens that in our data set longer genes have a slightly higher rate of synonymous substitutions and are not expressed in as broad a range of tissues. Therefore, plotting mean MCB for the randomized genes against breadth of expression for the real gene, we still find a weak positive correlation of the order of magnitude reported for the real genes ($P < 0.001$, $r^2 = 4.0\%$). Likewise we find in the mean MCB *vs.* K_s regression a weak negative correlation of about the order reported for the real genes [Li93, $P = 0.001$, $r^2 = 0.6\%$; Tamura and Nei method (TN93), $P = 0.002$, $r^2 = 0.5\%$]. Moreover, when we subtract the average bias of the random sequences from the bias of the real sequences, the correlation with breadth of expression disappears and with rates of substitution weakens considerably (expression, $P = 0.348$, $r^2 = 0.01\%$; K_s Li93, $P = 0.014$, $r^2 = 0.03\%$). Therefore, the most conservative interpretation of our data is that MCB does covary with expression breadth and K_s, but this is likely to be because of a tendency of larger genes having lower expression breadth and higher rates of silent site substitution. The data appear not to support the hypothesis that covariance is due to selection on codon usage *per se.* It should be noted that for 96% of the sequences the MCB value of the real data was higher than the mean value for the random sequences.

Dinucleotide effects and preferred codons: We are left trying to understand why there is such a large residual variance in codon usage after background nucleotide content is taken into account. One possibility is that the biases are caused by mutation biases or selection associated with particular dinucleotides. We performed a dinucleotide analysis on the whole sample (see METHODS) and also found that the sequences of the sample show significant biases in the appearance of dinucleotides from the expectations based on nucleotide content variations, consistent with previous observations (KARLIN and MRAZEK 1996). Dinucleotide bias explains part of the codon usage bias that we find in sequences in our sample when correcting for background nucleotide content. Most notably both TA and CG are avoided. It has been suggested that the dearth of CpG is probably related to the mutation of methylated CpG sites to TpG dinucleotides. By contrast, the dearth of TA may be owing to selection related to the susceptibility of UA in mRNA to RNase activity (BEUTLER *et al.* 1989; but see DURET and GALTIER 2000). A similar bias is also found in noncoding regions (KARLIN and MRAZEK 1996), suggesting some RNA-independent mechanism; however, the bias is significantly more profound in coding regions.

It is not the case, however, that dinucleotide effects can explain the totality of the residual bias. If there are significant biases that cannot be explained by dinucleotide effects, then this should be revealed by comparing the relative frequencies of the codons that encode amino acids with the same degree of degeneracy and that share the second nucleotide in their codons. So, for example, if dinucleotide biases explain codon usage bias, then the relative frequency of A-ending codons among the codons that specify glutamine (CAA, CAG) should be the same as the relative frequency of the A-ending co-

dons among the codons that specify glutamic acid (GAA, GAG).

For each gene, the relative frequencies for each codon were calculated with respect to the other codons that encode the same amino acid. Those amino acids whose codons have the same nucleotide at the second site and that have the same type of degeneracy were grouped. Three such groups can be formed: (1) tyrosine, histidine, asparagines, and aspartic acid; (2) glutamine, lysine, and glutamic acid of twofold degeneracy; and (3) proline, threonine, and alanine of fourfold degeneracy. Within each group of amino acids, subgroups of those codons that have the same nucleotides at the first and the second sites were formed. Within each subgroup the relative frequencies of codons were compared against each other with Mann-Whitney tests. A total of 21 comparisons were made (for amino acids with twofold degree of degeneracy, only one subgroup of codons was formed since the second subgroup is complementary). In this test, the CG content variations do not affect the comparisons because the relative frequencies for each amino acid are calculated with respect to the other codons that encode the same amino acid. Assuming that there are no diamino acid biases or other factors of bias than dinucleotide effects, then we can expect that the relative proportions of codons of different amino acids are nearly identical (*i.e.*, not significantly different) since they are expected to interact with similar proportions of nucleotides at the first position of the next codon. The major difference was found in the comparison of the codons CAA and GAA (encoding for glutamine and glutamic acid, respectively) with mean frequencies of 0.24 and 0.38, respectively. From all 21 possible comparisons within subgroups, only the comparison of the codons TAT and AAT (that encode for tyrosine and asparagine, respectively) was not significant with an α value of 0.05. All but 4 were significantly different with an α value of 0.01.

Some of the differences observed might be due to the existence of trinucleotide biases, diamino acid biases, or any more elaborated mutation patterns. These results show, however, that dinucleotide effects cannot alone account for all of the observed distribution of codons in human genes.

DISCUSSION

We found that codon usage bias in mammalian genes is not completely explained by background nucleotide content variation. We therefore developed a method to study the influence of other variables on codon usage. Unlike other methods ours appears to be insensitive to influence from rare amino acids. When we apply this method to a sample of human sequences, correcting expected distributions for background nucleotide content, we find that codon usage bias covaries with breadth of expression and is inversely correlated with the rate of synonymous substitution. This could suggest selective pressures related to translation efficiency, as has been conjectured (Debry and Marzluff 1994). However, the fact that these two correlations disappear when the effect of gene length is included suggests that gene length could be a more relevant variable and that the suggestive results are just artifacts. It is nonetheless interesting to find a weak tendency for short genes to have broader expression and lower synonymous substitution rates. These effects are, however, so weak that it may be improper to suppose that they have any biological meaning.

We also observe that there are codons that are consistently over- or underrepresented. This pattern can be explained in part by dinucleotide biases that also influence codon usage. However, we have also shown that not all the bias can be explained by such a simple mutational bias. While the cause of the remaining bias is uncertain, we fail to provide support for the hypothesis that codon usage is owing to selection.

Can we be sure that selection does not affect codon usage in mammals? While the above results would tend to suggest an absence of selection, as might be assumed to be the dominant position, several caveats must be noted. First, the dearth of TA dinucleotides may be a result of selection, as we discussed. However, Duret and Galtier (2000) argue that this is a methodological artifact. Second, our expression analysis looked at breadth of expression, not rate of expression. Nonetheless, the breadth of expression is known to covary with the intensity of purifying selection acting on the nonsynonymous sites (Duret and Mouchiroud 2000), so may reasonably be taken as a covariate to the strength of purifying selection.

Third, we need to understand how to resolve the present findings with the result that there are dramatic increases in the amount of gene expression observed when foreign sequences, to be expressed in mammalian cells, are modified to avoid having rare codons. One possibility is that negative results are not reported and therefore we are left only with the cases in which the increase in expression could be due to the change in some synonymous sites rather than the effect of codon usage *per se*. On the other hand, this observation could indeed be indicative of selective pressures related to translation efficiency acting on codon usage distributions. However, because we are using a method that measures distance from random use, rather than the degree in which optimal codons are used, we might not have adequate resolution to detect the patterns. Using a method to measure codon bias based on the degree of use of optimal codons, but correcting for the background nucleotide bias, could allow recovery of evidence of weak selective pressures acting on coding sequences in mammals. In the meantime, we may conclude that codon usage bias covaries with expression breadth and the rate

of synonymous evolution in humans but that this is not evidence for selection.

We thank Laurent Duret, Brian Charlesworth, Jody Hey, and two anonymous referees for comments on an earlier version of the manuscript. This work was funded by a grant from Conacyt to A.O.U. and by the Biotechnology and Biological Sciences Research Council (BBSRC) and the Royal Society to L.D.H.

LITERATURE CITED

BERNARDI, G., 1995 The human genome: organization and evolutionary history. Annu. Rev. Genet. **29:** 445–476.

BERNARDI, G., D. MOUCHIROUD and C. GAUTIER, 1997 Isochores and synonymous substitutions in mammalian genes, pp. 197–208 in *DNA and Protein Sequence Analysis*, edited by M. J. BISHOP and C. J. RAWLINGS. IRL Press, Oxford.

BEUTLER, E., T. GELBART, J. H. HAN, J. A. KOZIOL and B. BEUTLER, 1989 Evolution of the genome and the genetic-code—selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA **86:** 192–196.

DEBRY, R. W., and W. F. MARZLUFF, 1994 Selection on silent sites in the rodent H3 histone gene family. Genetics **138:** 191–202.

DUNN, K. A., J. P. BIELAWSKI and Z. H. YANG, 2001 Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157:** 295–305.

DURET, L., and N. GALTIER, 2000 The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. Mol. Biol. Evol. **17:** 1620–1625.

DURET, L., and L. D. HURST, 2001 The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. **18:** 757–762.

DURET, L., and D. MOUCHIROUD, 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

DURET, L., and D. MOUCHIROUD, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17:** 68–74.

EYRE-WALKER, A. C., 1991 An analysis of codon usage in mammals—selection or mutation bias. J. Mol. Evol. **33:** 442–449.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria—correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GOUY, M., F. MILLERET, C. MUGNIER, M. JACOBZONE and C. GAUTIER, 1984 Acnuc—a nucleic-acid sequence data-base and analysis system. Nucleic Acids Res. **12:** 121–127.

HANAI, R., and A. WADA, 1988 The effects of guanine and cytosine variation on dinucleotide frequency and amino-acid composition in the human genome. J. Mol. Evol. **27:** 321–325.

IKEMURA, T., and K. WADA, 1991 Evident diversity of codon usage patterns of human genes with respect to chromosome-banding patterns and chromosome-numbers—relation between nucleotide-sequence data and cytogenetic data. Nucleic Acids Res. **19:** 4333–4339.

KANAYA, S., Y. YAMADA, Y. KUDO and T. IKEMURA, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238:** 143–155.

KARLIN, S., and J. MRAZEK, 1996 What drives codon choices in human genes? J. Mol. Biol. **262:** 459–472.

KING, J. L., and T. H. JUKES, 1969 Non-Darwinian evolution. Science **164:** 788–798.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

LEVY, J. P., R. R. MULDOON, S. ZOLOTUKHIN and C. J. LINK, 1996 Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. Nat. Biotechnol. **14:** 610–614.

LI, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98:** 5688–5692.

MORIYAMA, E. N., and J. R. POWELL, 1997 Codon usage bias and tRNA abundance in Drosophila. J. Mol. Evol. **45:** 514–523.

POWELL, J. R., and E. N. MORIYAMA, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

SHARP, P. M., and W. H. LI, 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. **4:** 222–230.

SHARP, P. M., T. M. F. TUOHY and K. R. MOSURSKI, 1986 Codon usage in yeast—cluster-analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5143.

SHARP, P. M., M. AVEROF, A. T. LLOYD, G. MATASSI and J. F. PEDEN, 1995 DNA-sequence evolution—the sounds of silence. Philos. Trans. R. Soc. Lond. Ser. B **349:** 241–247.

STENICO, M., A. T. LLOYD and P. M. SHARP, 1994 Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res. **22:** 2437–2446.

WELLS, K. D., J. A. FOSTER, K. MOORE, V. G. PURSEL and R. J. WALL, 1999 Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch. Transgenic Res. **8:** 371–381.

WRIGHT, F., 1990 The effective number of codons used in a gene. Gene **87:** 23–29.

ZHOU, J., W. J. LIU, S. W. PENG, X. Y. SUN and I. FRAZER, 1999 Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. J. Virol. **73:** 4972–4982.

ZOLOTUKHIN, S., M. POTTER, W. W. HAUSWIRTH, J. GUY and N. MUZYCZKA, 1996 A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. J. Virol. **70:** 4646–4654.

Communicating editor: J. HEY

Chapter three

Sequence Characteristics and Expression Patterns in Human Genes

Letter

The Signature of Selection Mediated by Expression on Human Genes

Araxi O. Urrutia and Laurence D. Hurst[1]

Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK

As the efficacy of natural selection is expected to be a function of population size, in humans it is usually presumed that selection is a weak force and hence that gene characteristics are mostly determined by stochastic forces. In contrast, in species with large population sizes, selection is expected to be a much more effective force. Evidence for this has come from examining how genic parameters vary with expression level, which appears to determine many of a gene's features, such as codon bias, amino acid composition, and size. However, not until now has it been possible to examine whether human genes show the signature of selection mediated by expression level. Here, then, to investigate this issue, we gathered expression data for >10,000 human genes from public data sets obtained by different technologies (SAGE and high-density oligonucleotide chip arrays) and compared them with gene parameters. We find that, even after controlling for regional effects, highly expressed genes code for smaller proteins, have less intronic DNA, and higher codon and amino acid biases. We conclude that, contrary to the usual supposition, human genes show signatures consistent with selection mediated by expression level.

It is usually assumed that in humans, gene characteristics such as gene length and amino acid composition are mostly determined by stochastic processes (Eyre-Walker 1991; Sharp et al. 1995; Smith and Hurst 1999). The only sources of significant selective pressure would be those related to protein function optimization. Because protein synthesis has an associated cost to the cell, selection should favor changes in gene sequences that make protein synthesis more efficient or reduce its costs. The strength of selection related to protein synthesis efficiency should be higher for those genes transcribed in large quantities. Observations from several unicellular and invertebrate species have shown that expression profiles of genes covary with a variety of sequence parameters (Akashi 2001) such as gene length (Coghlan and Wolfe 2000; Jansen and Gerstein 2000), codon usage bias (Gouy and Gautier 1982; Sharp et al. 1986; Duret and Mouchiroud 1999; Coghlan and Wolfe 2000), and amino acid composition (Jansen and Gerstein 2000; Akashi and Gojobori 2002). These patterns have been interpreted as evidence of selection acting to increase protein synthesis efficiency and to reduce associated costs.

In human and other mammalian species, it has been suggested that gene sequences should not show the effects of natural selection to increase protein synthesis efficiency because of their small population sizes. Therefore, no relationship between expression and gene characters is expected (Eyre-Walker 1991; Sharp et al. 1995; Smith and Hurst 1999). Some evidence of selection acting on codon usage in mammalian genes has been reported in the past, but these studies are based on samples of limited size and/or do not test directly whether codon usage is related to activity levels of genes (Eyre-Walker 1991; Debry and Marzluff 1994; Iida and Akashi 2000). Recently, it was shown (Castillo-Davis et al. 2002) that expression patterns are related to intron sizes in human genes. However, this study does not take into account the possible influence of regional mutational biases influencing the local level of insertions and deletions. In addition, some reservations should be taken when using data derived from EST libraries used in this study to estimate levels of activity. Here, using two independent data sets of gene expression and correcting for regional effects, we provide a systematic analysis to clarify whether human genes show signatures consistent with expression-mediated selection.

If selection is acting on gene sequences, then we expect them to be modified to maximize expression efficiency. This effect should be particularly pronounced for highly expressed genes. To address this issue, we compared estimates of expression against gene characteristics. For this purpose, we assembled expression data from publicly available SAGE libraries from NCBI collected at different laboratories and representing 22 different tissues (see Methods). Serial Analysis of Gene Expression technology (SAGE) allows the measurement of expression profiles for large numbers of genes in a relatively unbiased way by avoiding gene-specific mRNA screening (Velculescu et al. 1995). In addition, we also used the comprehensive analysis of gene expression data using high-density oligonucleotide array technology recently released and representing 29 different tissues (Su et al. 2002; see Methods). Because in this data set all tissues were tested for the same genes, there is no sampling bias caused by the screening for different sets of genes in different tissues.

RESULTS

Transcription–Translation Efficiency and Gene Expression

Does selection act on coding sequences of genes to maximize translation and/or transcription efficiency and gene position? If so, then we may expect highly expressed genes to produce shorter proteins to reduce translation costs. This is what we find: Genes of higher expression produce only short proteins, and we find a significant negative correlation between protein length and mean level of expression ($R = 0.182$, $p < 0.0001$; $N = 8212$; see Table 1). Similarly, if transcription is costly, we might expect selection to act on intron length (Hurst et al. 1996). We found that, indeed, highly expressed genes have reduced total intron content ($R = 0.181$, $p < 0.0001$; $N = 7967$; Fig. 1). We found that intron and protein length are correlated. To examine whether both exon and intron lengths are independently related to expression levels, we performed multiple regression analyses correcting for intron and protein length, respectively. Expression levels are significantly related to intron and protein lengths after correction ($\beta = -0.221$, $p < 0.0001$; $\beta = -0.100$, $p < 0.0001$).

[1]**Corresponding author.**
E-MAIL l.d.hurst@bath.ac.uk; FAX 44-1225-826779.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.641103. Article published online before print in September 2003.

13:2260–2264 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

Table 1. Results From Multiple-Regression Analyses of Level of Gene Expression and Length With Gene Parameters When Including GC3 Content

Dependent variable	Database	Pearson correlation with level of expression ($p < 0.0001$)	Effect of expression when controlling for regional effects ($p < 0.0001$)[a]
Protein length[b]	SAGE	$R = 0.182$	$\beta = -0.175$
	Chip Array	$R = 0.194$	$\beta = -0.200$
Intron length[b]	SAGE	$R = 0.181$	$\beta = -0.403$
	Chip Array	$R = 0.198$	$\beta = -0.369$
Codon bias (MCB)[b]	SAGE	$R = 0.122$	$\beta = 0.019$
	Chip Array	$R = 0.180$	$\beta = 0.032$
AA complexity[b]	SAGE	$R = 0.062$	$\beta = -0.006$
	Chip Array	$R = 0.045$	NS

[a]Regional effects are gene density (average intergenic distance of two adjacent genes), base composition (intergenic base composition of 5000 bp at either side of gene), and recombination rate (average of recombination rate of nearest markers weighted by distance).
[b]Variables log transformed for analysis.

The compact nature of highly expressed genes is then consistent with the activity of selection. If selection has acted to maximize the efficiency of translation (as suggested by the correlation with protein size), we might also expect patterns of gene expression to be related to codon bias, as they are in several unicellular and invertebrate species (Gouy and Gautier 1982; Sharp et al. 1986; Duret and Mouchiroud 1999). In these species, certain tRNAs are more abundant than others, and selection favors, in highly expressed genes, the codons that match the most abundant isoacceptor (Sprinzl et al. 1996) or the most accurate one (Dix and Thompson 1989; Akashi 1994), thus resulting in a correlation between codon bias and expression level.

In mammals, evidence of codon usage bias and its possible relation to expression profiles has remained scarce. In these species, there is a great degree of heterogeneity in base composition along the genome (Bernardi 1995), and codon usage bias in mammalian genes has been interpreted as the result of regional base composition variations (Eyre-Walker 1991; Sharp et al. 1995; Smith and Hurst 1999). Nevertheless, some previous studies indicate that codon bias might be related to expression profiles. Two studies, one using histone genes, which are highly expressed (Debry and Marzluff 1994), and a second comparing codon preferences of alternatively spliced and constitutive exons (Iida and Akashi 2000), conclude that codon choice in highly expressed genes/constitutive exons deviates from the expected distribution, from flanking regions/alternatively spliced exons, respectively. Further support for a possible relationship between codon usage and expression levels of genes comes from studies in which the expression of nonmammalian genes in mammalian cells has been dramatically increased by the replacement of rare codons in the mammalian genome with common ones. This method of "codon optimization" has been used to increase expression of several genes (Levy et al. 1996; Wells et al. 1999; Zhou et al. 1999). All of these studies used a very limited sample size, and therefore their findings cannot be generalized to all mammalian genes.

Are the above isolated cases or is there is a broad relationship between expression and codon choice? In a previous study using a sample size of >2000 human genes, we showed that for most of the genes, codon usage bias is significantly higher than expected from background nucleotide composition (Urrutia and Hurst 2001). We now examine whether this residual bias is related to expression levels. To test this, we compared expression levels with codon bias in our gene data set. We measured codon bias using the methods of KM (Karlin and Mrazek 1996) and MCB (Urrutia and Hurst 2001; see Methods). Unlike more conventional measures (e.g., ENC), these two methods attempt to correct for background nucleotide variation. MCB has the advantage over KM of being less biased by amino acid composition. When correcting for nucleotide bias, we found that codon bias is correlated with level of expression (for KM, $R = 0.130$, $p < 0.0001$; for MCB, $R = 0.122$, $p < 0.0001$; $N = 6071$; Fig. 2). In a previous study (Urrutia and Hurst 2001), we showed that the MCB method is biased by protein length. Therefore, we assessed the correlation of expression level and codon bias after correction by length of protein. The correlation of MCB index with expression level remains significant after correcting for length of protein ($\beta = 0.048$, $p < 0.0001$).

Protein Synthesis and Expression Rates

Because of differences in the costs and biochemical properties associated with amino acid biosynthesis and/or with acquisition through the diet, we might expect genes expressed in large quantities to have a biased amino acid usage from that expected by their base composition. Evidence for a relation between expression levels and amino acid biases has been reported for yeast and bacteria (Jansen and Gerstein 2000; Akashi and Gojobori 2002). We examined the amino acid composition of genes and its relation to expression patterns. We observed a significant relation between amino acid usage and expression level for 16 out of 20 amino acids (after Bonferroni correction; see Table 2). However, because amino acid composition is also affected by background GC content (Singer and Hickey 2000), we corrected for the effect of GC3 content. All relationships remained significant even after correcting for gene length and GC3 content (after Bonferroni correction; see Table 2). It may be expected that the bias in the use of amino acids that we found would correspond to the avoidance of expensive to produce or scarce amino acids. Dufton (1997) developed an index of amino acid size/complexity based on the molecular weight and the shape of amino acids. We used this index as an indirect estimate of amino acid cost and examined its relationship to expression level. We find that, indeed,

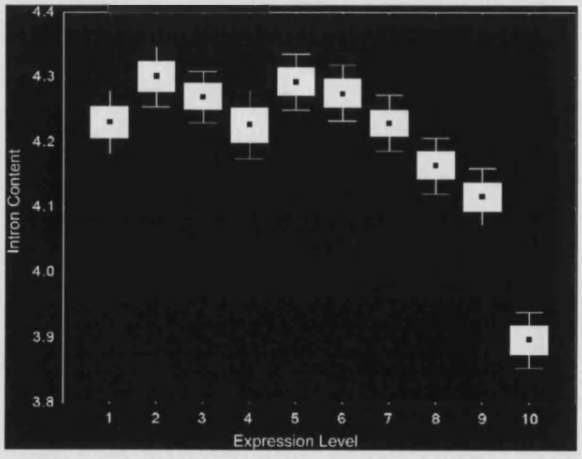


Figure 1 Intron content and expression level in human genes. Genes were split into 10 groups of an equal number of cases according to expression level. White dots represent the mean expression value for each group. Black boxes and error bars show the standard error with 68% and 95% of confidence.

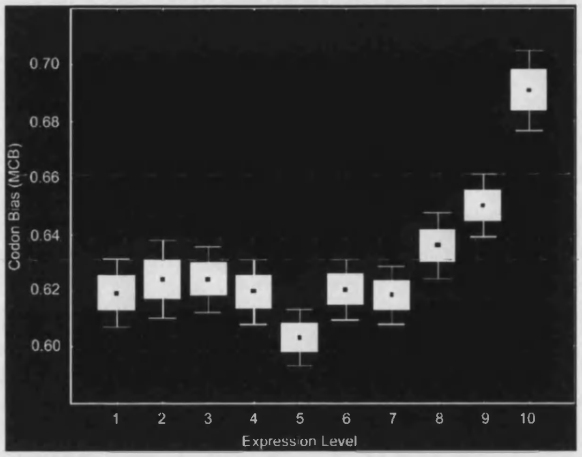


Figure 2 Codon bias (MCB) and level of expression. Codon usage bias MCB after correcting for background nucleotide content. Genes were split into 10 groups of an equal number of cases according to expression level. White dots represent the mean expression value for each group. Black boxes and error bars show the standard error with 68% and 95% of confidence.

there is a tendency to avoid the use of complex amino acids in highly expressed genes ($R = 0.062$, $p < 0.0001$; $N = 6223$; Table 1). More accurate estimates of the true cost of amino acid synthesis/acquisition for mammals would allow us to resolve the extent of the relationship between expression and amino acid choice.

Gene Position and Density Are Related to Expression Level

It has been previously reported that highly expressed genes tend to cluster in the human genome (Caron et al. 2001; Lercher et al. 2002). We confirm this: When we compared expression patterns of pairs of adjacent genes, we found significant similarity in expression level ($R = 0.09$, $p < 0.001$; $N = 4376$; see Methods). Note that all pairs of duplicated genes were removed (see Methods). We found that intergene spacers tend to be shorter for highly expressed genes ($R = 0.029$, $p < 0.0001$; $N = 8076$). This may possibly reflect an adaptation for more efficient gene transcription, but might alternatively reflect some regional mutational bias that tends to compact sequences in these regions, or differences in recombination rates (Hey and Kliman 2002). In addition, intron and protein lengths are correlated to intergenic distances (data not shown). Therefore, it is necessary to ask whether, controlling for regional effects, there remains a significant relationship between both intronic and protein sizes and expression level. On a multiple regression test, expression level is a relevant predictor of both protein and intron lengths after correcting them for intergenic length (see Table 1).

We have recently reported that highly expressed genes tend to be in GC-rich regions of the genome (Lercher et al. 2002); consistent with this, we found that highly expressed genes tend to have a higher GC content in the adjacent intergenic regions ($R = 0.065$, $p < 0.0001$; $N = 6430$; see Methods). Expression data derived from SAGE technologies could overestimate expression measures for GC-rich genes (Margulies et al. 2001). However, we find a similar pattern with chip-array-technology-derived data, for which no systematic errors have been reported, indicating that the relationship between expression level and GC content is not an artifact. We assessed, nevertheless, the relationship between expression level and protein and intron lengths, controlling for GC effects. The results of multiple regression analysis, however, show that level of expression is a relevant parameter for the above-discussed gene characteristics after correcting for GC content (see Table 1). Similar results were obtained correcting for GC3s (data not shown).

The control for GC content, in addition, in part controls for ancestral recombination rates (Marais 2003). But we also corrected our results using present estimates of recombination rates (Kong et al. 2002). There is a weak tendency for highly expressed genes to be situated in regions of higher recombination ($r = 0.013$, $p < 0.0001$, $N = 7987$). Expression level remains a relevant predictor of gene parameters after incorporation of recombination rate in the multiple regression analysis (see Table 1). However, as noted (Marais 2003), the present measures are both noisy and may well have little correlation to ancestral recombination rates; hence, interpretation of the above results from the best direct estimates must be limited.

DISCUSSION

Here we have evaluated the interaction between expression level of human genes and gene sequence characteristics. In sum, we find that highly expressed genes code for small proteins, have little intronic content, high codon bias, and tend to encode cheaper amino acids. These signatures found in human genes are consistent with the action of selective pressures to maximize protein synthesis efficiency in highly expressed genes.

In addition, we confirmed previous results on gene sorting by expression patterns and the relationship between expression patterns and GC content. We performed multiple regression analysis to rule out the possibility that these regional characters could potentially explain the relationship between expression patterns and gene characteristics. The relationship between expression level and intron and protein size can only in part be accounted for by regional compaction effects. Biases in codon and amino acid usage are not accounted for by GC bias or gene size. The relationship between expression rates and amino acid

Table 2. Amino Acid Usage and Expression Level (SAGE), Multiple Regression Analysis

Amino acid	One letter code	Pearson correlation with expression ($p < 0.0001$)	Effect of expression when controlling for GC3 content and gene length ($p < 0.0001$)
Alanine	A	0.100	$\beta = 0.336$
Arginine	R	NS	—
Aspargine	N	NS	—
Aspartic acid	D	0.100	$\beta = 0.562$
Cysteine	C	−0.077	$\beta = -0.458$
Glutamine	Q	−0.071	$\beta = -0.256$
Glutamic acid	E	0.055	$\beta = 0.561$
Glycine	G	0.084	$\beta = 0.418$
Histidine	H	−0.118	$\beta = -0.349$
Isoleucine	I	NS	—
Leucine	L	−0.105	$\beta = -0.888$
Lysine	K	0.126	$\beta = 1.049$
Methionine	M	0.071	$\beta = -0.142$
Phenylalanine	F	−0.032	$\beta = -0.155$
Proline	P	−0.045	$\beta = -0.401$
Serine	S	−0.145	$\beta = 0.785$
Threonine	T	0.045	$\beta = -0.086$
Tryptophan	W	−0.071	$\beta = -0.211$
Tyrosine	Y	NS	—
Valine	V	0.055	$\beta = 0.227$

Threshold of significance is defined after Bonferroni correction.

composition could be partly due to functional properties of the proteins associated with different expression levels.

Although the effects presented here are weak, it should be noted that similar results were obtained with two independent databases of gene expression obtained with different methodologies. In addition, in doing this work we have assumed a conservative approach when correcting all results for intergene spacers and GC content corrections not done in previous analysis (Castillo-Davis et al. 2002). The compaction of intergenic regions of highly expressed genes, however, need not reflect a mutation bias, but selective forces directly or indirectly related to expression patterns. In addition, codon bias has been estimated taking out any compositional biases, but these themselves could be partly driven by selection. Moreover, the correspondence between libraries representing the same tissue obtained with the same method is usually high ($r > 0.80$), whereas the correspondence of data obtained with different methods is low ($r < 0.60$; data not shown). These discrepancies are likely to add noise to our analyses and possibly derive from errors in the correspondence between oligonucleotides and/or tags and the gene represented. This should not affect our conclusions.

The results presented on gene length and expression patterns are consistent with those obtained in other multicellular eukaryotes but differ from observations in unicellular eukaryotes and bacteria, in which intron (Vinogradov 2001) and protein sizes (Moriyama and Powell 1998) are positively correlated to expression estimates. The patterns in unicellular organisms might be caused by increased expression gained by the inclusion of functional elements important for transcription regulation or splicing efficiency. The reversal of this along the evolutionary scale could be explained by the increased gene and genome size where most intergenic and intron sequences do not possess a function in gene regulation.

Where do our results leave the usual supposition that human population sizes are too small for selection to affect the properties that we have analyzed? Our results are probably largely in agreement with this general position. It is most notable that many of the results that we describe are not strong effects and in many cases appear to affect only the most highly expressed set of genes. We can imagine two reasons for this. First, only in this subset of genes is selection strong enough to have an appreciable effect. Classical theory postulates that for deleterious mutations not to be deterministically eliminated by selection, the selective coefficient, s, must be less than $1/2N_e$. Should these mutations not be eliminated, they would lead to genes tending to move away from optimal structure and codon usage. In the human genome, there may well be more mutations that are effectively neutral than in flies (humans having a smaller N_e), there nonetheless remains a respectable number of genes (the most highly expressed) in which $s \gg 1/2N_e$ for many mutations affecting level of expression. Second, and not mutually exclusively, we may be witnessing, in some part, the decay of selected features. As many of the features concerned may take time to reach equilibrium, we would expect that the most highly expressed genes would still retain many of the features of the prior action of selection. Analysis of the population genetics of insertion mutations in introns in highly expressed genes would then be interesting as the former model predicts that they may still be under counter selection, whereas the latter predicts that they may instead be effectively neutral.

METHODS

Sequence Information

Sequence information was obtained for genes from human genome annotations from build 30 of the NCBI site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). Although 26,297 genes were considered for the analysis, data for each parameter were not obtained for all of the genes; therefore, the actual number of genes used in comparisons varied as indicated in the text. Nucleotide sequences were retrieved from the Fa file from the NCBI site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/). Nucleotide composition was determined, and KM (Karlin and Mrazek 1996) and MCB values of codon bias were obtained after nucleotide corrections were obtained according to methods stated elsewhere (Urrutia and Hurst 2001). Nucleotide expectations for codon usage were based on the coding sequence of each gene and obtained according to Urrutia and Hurst (2001). This is preferable to using noncoding regions as these typically contain repetitive elements, regulatory sequences, and even RNA genes that would bias base composition. In all cases in which more than one alternative transcript was available, the largest was analyzed. Incomplete sequences were removed from analysis.

Intron–Exon Boundaries

Intron–exon boundaries, intergenic length, and the identity of neighboring genes were established from contig annotations from the human genome sequence (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). All overlapping genes were removed. Because contigs are not always continuous, adjacent genes were not determined for genes that were either the first or last genes within their contig.

Intergenic GC Content

The intergenic GC content was obtained from masked chromosome assembly in fasta format of build 30 at the NCBI site (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). A section adjacent to each side of each gene of a minimum 500 bp and a maximum of 5000 bp was used to estimate intergenic GC content. The GC content was not calculated for all overlapping genes. The analyses presented here refer to global GC content percentage, but similar results were obtained with the GC content of nonrepetitive sequences only or GC3s from coding regions of genes.

Duplicated Genes

Duplicated genes were removed from the analysis of adjacent genes. All genes were blasted against the two adjacent genes using the BLASTN downloadable version from the NCBI site (http://www.ncbi.nlm.nih.gov/BLAST/). All pairs of adjacent genes with an expected value of sequence similarity <0.01 were removed from analysis. The correlation coefficient was obtained from the comparison of rates of expression of adjacent genes, in which the order of the genes of each pair was randomly assigned. The correlation coefficient shown for expression rates of adjacent genes refers to the mean value of 100 of such correlations.

Recombination Data

Recombination data were obtained from Kong et al. (2002). The recombination rate indexes for each gene were derived from composing the recombination rates of the nearest marker at each side. The relative weight for each adjacent marker was determined by the distance separating the gene from the marker at each side.

Expression Data

SAGE expression data were collected from the NCBI site (http://www.ncbi.nlm.nih.gov/SAGE/). Only tags that matched to a single gene were taken into account. In addition, because tags were matched against reported sequences in GenBank and only a small percentage of these sequences contain a poly(A), tags containing poly(A) tails would only be matched against a small subset of the sequences. Therefore, all tags that ended in a stop codon followed by more than five As were discarded. All genes for which only one tag was detected in all libraries were also eliminated from the analysis as they potentially represent a sequencing error. Only libraries from normal tissues (noncancerous) were used in the study (43). Transcript counts for libraries corresponding to the same tissue were joined, and tags per mil-

lion were then calculated for each gene. The data on 8220 genes for 22 tissues were taken into account: brain, cerebellum, spinal cord, skin, vascular, T-cells, lymphocyte, muscle, retina, cornea, mammary glands, heart, lung, kidney, stomach, liver, pancreas, colon, peritoneum, uterus, ovary, and prostate. Those corresponding to the same tissue were averaged before obtaining a global measure of expression level.

High-density oligonucleotide array data was collected from the gene expression atlas site (http://expression.gnf.org). For any gene to be counted as expressed in a given tissue, a cutoff value on the expression index of 20 was defined. The data for 101 samples were available, corresponding to 28 noncancerous tissues: cerebellum, brain, cerebral cortex, caudate nucleus, amygdale, thalamus, corpus callosum, spinal cord, whole blood, testis, pancreas, placenta, pituitary gland, thyroid, prostate, ovary, uterus, dorsal root ganglia, salivary gland, trachea, lung, thymus, spleen, adrenal gland, kidney, liver, heart, umbilical vein, and endothelial cells.

From SAGE and chip array data, we could define two measures of level of expression: Peak expression, which is the highest value of expression of a gene in any tissue, and mean level of expression, the mean quantity of expression of a gene in all tissues in which it is expressed (if divided among all tissues, then this measure would be dependent on breadth of expression). As these two measures proved to be highly correlated ($R = 0.99$; data not shown), only mean expression is referred to here as level of expression. However, the results presented also apply to peak of expression of genes (data not shown).

The figures presented here refer to the analysis of SAGE data; similar results were also obtained when using chip-array data unless otherwise indicated in text and tables. Similar data are obtained when only genes not known to undergo alternative splicing are taken into account (data not shown). Indexes of expression level and lengths of coding and noncoding regions were log-transformed prior to analyses.

ACKNOWLEDGMENTS

We thank Brian Charlesworth, Laurent Duret, Hiroshi Akashi, and four anonymous referees for their helpful comments. We thank the BBSRC (L.D.H.) and the CONACyT and ORS (A.O.U.) for funding.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*—Natural selection and translational accuracy. *Genetics* **136:** 927–935.

———. 2001. Gene expression and molecular evolution. *Curr. Opin. Gene Dev.* **11:** 660–666.

Akashi, H. and Gojobori, T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc. Natl. Acad. Sci.* **99:** 3695–3700.

Bernardi, G. 1995. The human genome: Organization and evolutionary history. *Ann. Rev. Genet.* **29:** 445–476.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. 2002. Selection for short introns in highly expressed genes. *Nat. Genet.* **31:** 415–418.

Coghlan, A. and Wolfe, K.H. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16:** 1131–1145.

Debry, R.W. and Marzluff, W.F. 1994. Selection on silent sites in the rodent H3 histone gene family. *Genetics* **138:** 191–202.

Dix, D.B. and Thompson, R.C. 1989. Codon choice and gene-expression—Synonymous codons differ in translational accuracy. *Proc. Natl. Acad. Sci.* **86:** 6888–6892.

Dufton, M.J. 1997. Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? *J. Theor. Biol.* **187:** 165–173.

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila, Arabidopsis*. *Proc. Natl. Acad. Sci.* **96:** 4482–4487.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: Selection or mutation bias? *J. Mol. Evol.* **33:** 442–449.

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria—Correlation with gene expressivity. *Nucleic Acids Res.* **10:** 7055–7074.

Hey, J. and Kliman, R.M. 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics* **160:** 595–608.

Hurst, L.D., McVean, G.T., and Moore, T. 1996. Imprinted genes have few and small introns. *Nat. Genet.* **12:** 234–237.

Iida, K. and Akashi, H. 2000. A test of translational selection at 'silent' sites in the human genome: Base composition comparisons in alternatively spliced genes. *Gene* **261:** 93–105.

Jansen, R. and Gerstein, M. 2000. Analysis of the yeast transcriptome with structural and functional categories: Characterizing highly expressed proteins. *Nucleic Acids Res.* **28:** 1481–1488.

Karlin, S. and Mrazek, J. 1996. What drives codon choices in human genes? *J. Mol. Biol.* **262:** 459–472.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.

Levy, J.P., Muldoon, R.R., Zolotukhin, S., and Link, C.J. 1996. Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. *Nat. Biotech.* **14:** 610–614.

Marais, G. 2003. Biased gene conversion: Implications for genome and sex evolution. *Trends Genet.* **19:** 330–338.

Margulies, E., Kardia, S., and Innis, J. 2001. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.***29:** e60.

Moriyama, E.N. and Powell, J.R. 1998. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* **26:** 3188–3193.

Sharp, P.M., Tuohy, T.M.F., and Mosurski, K.R. 1986. Codon usage in yeast—Cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14:** 5125–5143.

Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G., and Peden, J.F. 1995. DNA-sequence evolution—The sounds of silence. *Phil. Trans. R Soc. Lond. B* **349:** 241–247.

Singer, G.A.C. and Hickey, D.A. 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.* **17:** 1581–1588.

Smith, N.G.C. and Hurst, L.D. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153:** 1395–1402.

Sprinzl, M., Steegborn, C., Hubel, F., and Steinberg, S. 1996. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **24:** 68–72.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.

Urrutia, A.O. and Hurst, L.D. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159:** 1191–1199.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene-expression. *Science* **270:** 484–487.

Vinogradov, A.E. 2001. Intron length and codon usage. *J. Mol. Evol.* **52:** 2–5.

Wells, K.D., Foster, J.A., Moore, K., Pursel, V.G., and Wall, R.J. 1999. Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch. *Transg. Res.* **8:** 371–381.

Zhou, J., Liu, W.J., Peng, S.W., Sun, X.Y., and Frazer, I. 1999. Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *J. Virol.* **73:** 4972–4982.

WEB SITE REFERENCES

ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/; Chromosome annotations, human genome, NCBI.

ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/; RNA file, human genome, NCBI.

http://expression.gnf.org; Expression Atlas.

http://www.ncbi.nlm.nih.gov/BLAST/; BLAST tools, NCBI.

http://www.ncbi.nlm.nih.gov/SAGE/; SAGE, NCBI.

Received July 17, 2002; accepted in revised form July 28, 2003.

Chapter Four

Gene Order and Gene Expression

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genetics 31: 180-183.

letter

Clustering of housekeeping genes provides a unified model of gene order in the human genome

Martin J. Lercher*, Araxi O. Urrutia* & Laurence D. Hurst

*These authors contributed equally to this work.

Published online: 6 May 2002, DOI: 10.1038/ng887

It is often supposed that, except for tandem duplicates, genes are randomly distributed throughout the human genome. However, recent analyses suggest that when all the genes expressed in a given tissue (notably placenta[1] and skeletal muscle[2]) are examined, these genes do not map to random locations but instead resolve to clusters. We have asked three questions: (i) is this clustering true for most tissues, or are these the exceptions; (ii) is any clustering simply the result of the expression of tandem duplicates and (iii) how, if at all, does this relate to the observed clustering of genes with high expression rates[3]? We provide a unified model of gene clustering that explains the previous observations. We examined Serial Analysis of Gene Expression (SAGE)[4] data for 14 tissues and found significant clustering, in each tissue, that persists even after the removal of tandem duplicates. We confirmed clustering by analysis of independent expressed-sequence tag (EST) data. We then tested the possibility that the human genome is organized into subregions, each specializing in genes needed in a given tissue. By comparing genes expressed in different tissues, we show that this is not the case: those genes that seem to be tissue-specific in their expression do not, as a rule, cluster. We report that genes that are expressed in most tissues (housekeeping genes) show strong clustering. In addition, we show that the apparent clustering of genes with high expression rates[3] is a consequence of the clustering of housekeeping genes.

Tight clustering of co-expressed genes, most notably in operons, is common in prokaryotes. Genes that encode proteins that interact tend to be linked and to stay linked[5]. Operons have also been described in *Caenorhabditis elegans*[6], but in eukaryotes it is typically assumed that genes are randomly distributed. Nonetheless, reports have suggested that gene location might not be random[1–3,7,8]. A recent study showed that genes with high median rates of expression tend to cluster in the human genome[3]; however, this report incorporated data from several cancerous tissues, and did not control for the correlation between tandem duplicates. A systematic analysis of the clustering of comparably expressed human genes is thus currently lacking.

Table 1 • Correlations between breadth and peak rate of expression, GC content and gene density

		N	*r*	r^2
Breadth	gene density	11,612	0.042	0.002
Breadth	GC	11,549	0.167	0.028
Log(peak rate)	gene density	8,224	0.062	0.004
Log(peak rate)	GC	8,170	0.185	0.034
Log(peak rate)	breadth	8,224	0.693	0.480

All correlations are highly significant ($P < 10^{-5}$ from randomizations).

We obtained SAGE expression profiles for 11,612 human genes across 14 normal tissues (see Methods). We found a high correlation of breadth of expression and logarithm of peak expression rate ($r^2 = 0.48$, $P < 10^{-5}$). Although both breadth and peak rate are correlated with gene density and nucleotide composition (Table 1), these correlations are weak and do not account for the observed clusters. Both rate and breadth of expression show significant chromosomal heterogeneity, even when sex chromosomes are excluded ($P < 10^{-5}$ in each case; Table 2). The strongest outlier is the Y chromosome, with low mean breadth (0.7, compared with 3.8 for the whole genome), and low mean rate (log(peak rate) = 1.5, compared with 1.9 for the whole genome). There is also strong within-chromosome heterogeneity in expression breadth, consistent with clusters of housekeeping genes (as an example, see chromosome 11 in Fig. 1).

As genes duplicated in tandem may be similarly expressed for purely mechanistic reasons (such as *HOX* clusters and globin genes), we excluded such genes from further analyses (see Methods). This reduced the sample size to 5,112 genes with known positive expression in at least one normal tissue. To test for breadth-specific clustering, we subdivided the data into tissue-specific (breadth ≤ 2), intermediate and housekeeping (breadth ≥ 9) genes. For each class, we calculated the gene dispersion as the mean over variance of gene density, in sliding windows of 300 kb. Under a null model without clustering and with genes evenly distributed along chromosomes, gene counts per window should show a Poisson distribution, with an expected dispersion of 1. Owing to the uneven distribution of gene density along chromosomes, observed and expected dispersion of each class is below 1 (Table 3). The dispersion of housekeeping genes is significantly smaller than expected ($P < 10^{-5}$), whereas no such effect is seen for intermediate or tissue-specific genes (the apparent 'over-dispersion' of tissue-specific genes is due to randomization with clustered housekeeping genes). We confirmed this result for SAGE data with independent EST expression data for 6,298 genes in 60 tissues (see Methods). Housekeeping genes defined by EST (breadth ≥ 19) also show highly significant under-dispersion, whereas no such effect is seen for genes with low or intermediate expression breadth (Table 3).

Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK. Correspondence should be addressed to L.D.H. (e-mail: l.d.hurst@bath.ac.uk).

Table 2 • Chromosomal heterogeneity in expression breadth and rate

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
N	1227	726	649	416	540	664	514	382	442	434	690	644	202	378	349	483	684	184	789	329	135	286	439	26
Breadth mean	3.90	3.73	3.61	3.21	3.41	3.42	4.02	3.56	3.73	3.70	3.98	3.57	3.11	4.02	3.78	4.70	4.02	2.86	4.54	4.07	3.79	4.37	3.13	0.73
*P**	0.820	0.321	0.112	0.001	0.011	0.006	0.903	0.117	0.360	0.304	0.885	0.069	0.006	0.857	0.471	1	0.931	**0.0003**	1	0.890	0.492	0.992	**$<10^{-5}$**	**$<10^{-5}$**
N	863	528	454	278	357	437	365	257	324	319	491	441	150	285	272	391	489	119	568	244	92	219	270	11
Log (rate)mean	1.91	1.85	1.84	1.88	1.88	1.85	1.93	1.91	1.85	1.85	1.92	1.88	1.75	1.85	1.83	1.89	1.92	1.77	2.02	1.90	1.93	1.89	1.82	1.50
*P**	0.964	0.029	0.023	0.427	0.387	0.041	0.951	0.755	0.065	0.101	0.934	0.407	**$<10^{-5}$**	0.125	0.031	0.633	0.964	0.002	1	0.634	0.799	0.525	0.007	**0.0004**

**P* indicates the probability of finding a smaller mean value for a randomized genome. Significant values ($P < 0.025/24$ or $P > 1–0.025/24$, with Bonferroni correction) are shown in bold.

We then examined the peak rate of expression, the highest rate found for each gene across all tissues. We calculated the dispersion of genes with low (≤ 37 cpm), intermediate and high (≥ 134 cpm) peak rates of expression for the SAGE data, purged of tandem duplicates (Table 3). We found significant clustering of highly expressed genes, confirming that the previous observation[3] was not the result of similar expression of tandem duplicates. As expression rate and breadth are correlated (Table 1), the two clustering effects may not be independent. To test this, we repeated our dispersion calculation with an altered randomization protocol. When calculating 'random' dispersion values for housekeeping genes, we permuted only genes with similar expression rates. Although this reduces the difference between observed and expected dispersion (Table 3), housekeeping genes remain significantly under-dispersed. Conversely, when calculating 'random' dispersion values for highly expressed genes by permuting genes only in classes of similar expression breadth, the under-dispersion becomes nonsignificant. Thus, the clustering of genes with high expression rates is caused by the nonrandom distribution of housekeeping genes. When randomizing within classes of genes with similar nucleotide composition, or similar surrounding gene density, the significant under-dispersion of both housekeeping and highly expressed genes remains unchanged ($P < 10^{-5}$ in each case). Thus, neither composition nor gene density underlies the observed effects.

To examine the clustering of co-expressed genes, we calculated an index of co-expression (ICE) for the tandem duplicate–free SAGE and EST data sets (Fig. 2). The local similarity in expression is higher than expected by chance for all distances $d \leq 1$ Mb. It is significant for $d < 500$ kb in the SAGE data and $d < 300$ kb in the EST data. Although this measure can indicate the size of expression modules, it does not reveal how many clusters there are. We analyzed the distribution of physical cluster sizes from SAGE data, using a restrictive definition of clusters (see Methods). There is an excess of expression clusters over random expectations for cluster sizes up to approximately 350 kb (see Web Fig. A online). As an example, the largest such un-interrupted cluster is listed in Web Table A online. All genes in this cluster are expressed in at least seven tissues and can thus be classified as housekeeping genes. We therefore investigated whether the clustering of tissue-specific genes is an artifact of the clustering of housekeeping genes. We repeated the calculation of ICE for 'random' genomes, but this time permuting genes only within classes of similar expression breadth. In contrast to the results shown in Fig. 2, now only the nearest neighbors showed significant co-expression (SAGE: $d < 100$ kb, $P = 0.011$; EST: $d < 200$ kb, $P = 6 \times 10^{-5}$) . As EST data allows only an imprecise estimate of expression breadth (there are many false negatives), it is not unexpected to find stronger remnants of under-dispersion in these data.

Of all the genes expressed in any individual tissue, it seems that only a small minority are restricted to the tissue, whereas most carry out housekeeping functions[2]. Locating all genes expressed in a tissue thus reveals clustering of housekeeping genes. When studying the gene distribution for the tandem duplicate–free SAGE data, we found that all 14 tissues showed significant under-dispersion. When permuting genes only within classes of similar breadth, this under-dispersion becomes nonsignificant for all except one tissue (Table 4). A similar result is found for EST data: 40 of 60 tissues show significant under-dispersion; this is reduced to 17 tissues when randomizing genes within classes of similar breadth (data not shown). The less dramatic result of the EST data may be expected, as EST studies do not give an exhaustive overview of expression, and the available data contain many false negatives.

We have shown that housekeeping genes cluster and that this leads to both the clustering of highly expressed genes and the apparent clustering of all genes expressed in any one tissue. The selective pressure that promotes clustering is not simply related to the rate of transcription; it favors the linkage of the genes that

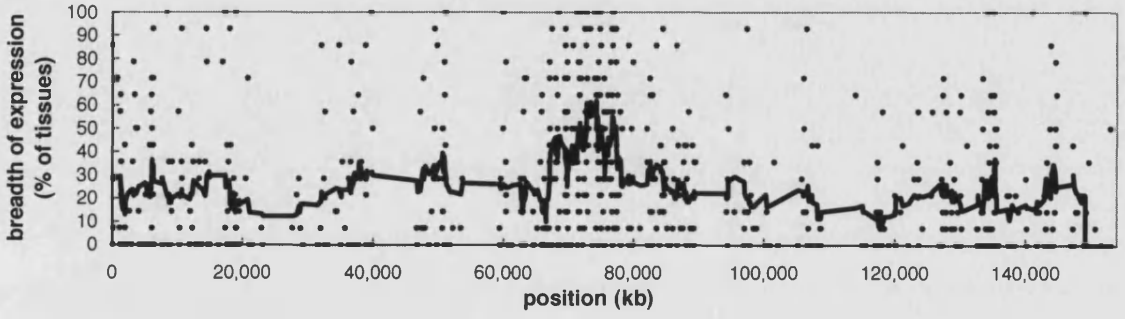


Fig. 1 Map of breadth of expression of the genes on chromosome 11. The solid line shows a sliding window average over 15 neighboring genes.

Table 3 • Dispersion of breadth- and rate-specific genes

		N	Dispersion	P*	Dispersion rand. disp.		P*	Dispersion rand. disp.
SAGE Breadth	low	1,656	0.946	1.00	1.095	randomization	0.998	1.033
	intermediate	2,332	0.808	0.20	0.988	in rate classes:	0.242	0.991
	high	1,120	0.796	$<10^{-5}$	0.881		0.00017	0.943
SAGE Peak rate	low	1,334	0.957	1.00	1.078	randomization	0.917	1.016
	intermediate	2,531	0.801	0.34	0.995	in breadth classes:	0.348	0.995
	high	1,242	0.805	$<10^{-5}$	0.901		0.082	0.978
EST Breadth	low	1,960	0.905	0.998	1.043		–	–
	intermediate	3,690	0.776	0.489	1.000		–	–
	high	1,927	0.777	$<10^{-5}$	0.894		–	–

*P indicates the probability of finding an equal or lower dispersion from 100,000 random permutations of gene positions.

are active across all tissues. Why do we see clustering of housekeeping genes? Higher-order chromatin structure[9], and thus accessibility of different genomic regions to the transcription machinery, may vary according to cell type. In this case, it might be advantageous to assemble housekeeping genes to some 'common ground' that remains in an open conformation across all cells. However, such a speculative link between expression and clustering requires analysis of chromosomal organization and gene control in the genomic regions concerned.

For the tissues examined here, we have shown that little clustering exists beyond that of housekeeping genes. However, we do not exclude the possibility of special cases of tissue-specific clustering. It has been suggested that clustering may be selectively favored because it allows linkage disequilibrium to more easily be maintained[10]; one example is the MHC[7,11]. Imprinting, such as that of placentally expressed genes, may provide another mechanism underlying tissue-specific clustering[12].

Methods

SAGE data. We used publicly available SAGE[4] data. We obtained a reliable mapping of UniGene[13] groups to *Nla*III SAGE tags from SAGEmap[14] at NCBI. Each UniGene group consists of all GenBank sequences representing the same human gene. Hereafter, each such group is referred to as a 'gene' and represented by its longest RefSeq sequence. Tags mapping to more than one gene were excluded. We located 11,612 RefSeq genes on the August 2001 Golden Path assembly of the human genome, each labeled unambiguously by at least one SAGE tag. This set of gene–tag combinations was cross-linked to the quantitative expression profiles at SAGEmap. We found that 8,367 genes showed positive expression in at least 1 of 35 libraries representing 14 normal tissues. If a tag had been counted only once in one tissue, this was most likely due to a sequencing error, and we discounted the observation. Adding all counts for libraries representing the same tissue type, we converted absolute tag counts to relative tag counts (cpm, counts per million). For each gene, we then calculated breadth of expression (number of tissues with positive expression). For those genes with positive expression in at least one tissue, we also calculated the peak rate of expression (maximum cpm). We sorted genes according to breadth, and independently according to rate, into eight classes of approximately equal size, respectively.

We obtained nucleotide composition (GC fraction) in contiguous 20-kb windows from the Golden Path assembly. We estimated gene density by counting the number of confirmed (RefSeq) genes in contiguous 300-kb windows on the Golden Path assembly. Genes were further sorted according to GC fraction and gene density into eight and seven classes of approximately equal size, respectively.

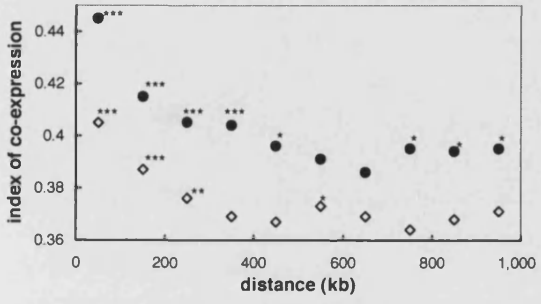


Fig. 2 Index of co-expression between genes, shown in a histogram in distance windows of 100 kb, for SAGE data (solid dots) and EST data (open diamonds). The significance level from randomizations is indicated with asterisks: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

EST data. Each UniGene[14] group contains not only the RefSeq coding sequence, but also all ESTs mapping to the gene. We used these ESTs to cross-link genes to EST libraries constructed from normal tissue samples, including only libraries containing at least 50 ESTs. This resulted in a data set of 11,382 genes, each known to be expressed in at least 1 of 60 normal tissues (13 prenatal and 47 postnatal). We calculated breadth of expression as the number of tissues with positive expression information; genes were sorted accordingly into eight classes of approximately equal size.

Removal of tandem duplicates. We developed a criterion to remove duplicated genes, as these are likely to have similar expression profiles resulting from their common history. From gene family trees in HOVERGEN[15], we selected sets of homologous human genes that diverged early in the evolution of vertebrates—that is, genes whose branches contained a high number of internal non-primate branches between them. This resulted in a set of 70 genes in 18 gene families. We carried out pair-wise, stand-alone protein BLAST (standard settings, word size 2 or 1) for 124 gene pairs within families and 306 gene pairs between different families. We found that 93% of gene pairs resulting from a duplication after the appearance of vertebrates had expect (E) values of less than 0.2, whereas 90% of gene pairs from different gene families had $E \geq 0.2$. We therefore used this value to define tandem duplicates. To remove tandem duplicates from our expression profile data set, we carried out pair-wise BLAST for all gene pairs within 1 Mb of each other and removed one gene of each pair having $E < 0.2$. This resulted in a duplicate-free data set of 5,112 genes known to be expressed in at least 1 of 14 normal tissues (SAGE), and a second data set of 6,298 genes known to be expressed in at least 1 of 60 normal tissues (EST).

Statistics. To assess correlations, we calculated Pearson's correlation coefficient (r) and compared this with 100,000 random data pairings. We assessed chromosomal heterogeneity of expression breadth with the test function,

$$\chi^2 = \sum_i \frac{(b_i - b)^2}{b} ,$$

with mean breadth b_i of the genes on chromosome i, and mean breadth b of all genes. This was compared with the corresponding values from 100,000 random genomes, each obtained by permuting the chromosomal positions of all genes. We estimated chromosomal heterogeneity of expression rate in a similar manner.

Table 4 • Dispersion for co-expressed genes in individual tissues from SAGE data

Tissue	N	dispersion	All genes P*	All genes dispersion / rand. disp.	Within breadth classes P*	Within breadth classes Dispersion / rand. disp.
White matter	3,016	0.730	$<10^{-5}$	0.941	0.77	1.008
Prostate	2,991	0.698	$<10^{-5}$	0.898	0.0088	0.977
Ovary epithelium	2,528	0.717	$<10^{-5}$	0.890	0.027	0.979
Vascular endothelium	2,506	0.751	$<10^{-5}$	0.930	0.987	1.024
Lung	2,254	0.791	0.0037	0.962	1.000	1.052
Mammary gland epithelium	2,144	0.750	$<10^{-5}$	0.904	0.39	0.997
Kidney (embryonic)	1,991	0.751	$<10^{-5}$	0.894	0.046	0.978
Colon epithelium	1,776	0.768	$<10^{-5}$	0.899	0.063	0.978
Pancreas epithelium	1,581	0.737	$<10^{-5}$	0.849	$<10^{-5}$	0.931
Astrocyte	1,567	0.796	$<10^{-5}$	0.917	0.799	1.012
Thalamus	1,349	0.806	$<10^{-5}$	0.910	0.16	0.984
Peritoneum	1,346	0.820	$<10^{-5}$	0.926	0.51	1.000
Kidney	933	0.835	$<10^{-5}$	0.911	0.11	0.979
Leukocyte	912	0.870	$<10^{-5}$	0.944	0.66	1.008

*P indicates the probability of finding an equal or lower dispersion from 100,000 random permutations of gene positions. All tissues show highly significant under-dispersion ($P < 0.05/14$, with Bonferroni correction); this becomes non-significant (except for pancreas) when comparing to randomizations where genes are permuted only within classes of similar expression breadth.

To determine whether genes cluster according to breadth of expression, we subdivided the data into genes with low expression breadth (SAGE: breadth ≤ 2, 32% of data; EST: breadth ≤ 5, 26% of data), intermediate expression breadth and high expression breadth (SAGE: breadth ≥ 9, 22% of data; EST: breadth ≥ 19, 25% of data). We then calculated the dispersion of the gene density for each of these classes: d = mean/variance of breadth-specific gene number in sliding windows of 300 kb width and a step size of 100 kb. To compare this with the results expected under a model in which gene order is random, we randomly permuted the gene positions of all genes 100,000 times and recalculated the breadth-specific dispersion. To test if any breadth-specific under-dispersion is a secondary effect caused by clustering according to rate of expression, we repeated this randomization procedure for the SAGE data, this time permuting gene positions only within each of the eight classes of similar expression rate. We used a corresponding protocol to test the SAGE data for clustering according to rate of expression. Genes were subdivided into classes of low expression rate (rate ≤ 37 cpm, 26% of data), intermediate expression rate and high expression rate (rate ≥ 134 cpm, 24% of data). This was compared with the results expected from a model with random gene order by permuting all gene positions 100,000 times. It was then also compared with the expected results from a model where genes are clustered according to breadth of expression, by permuting gene positions within only the eight classes of similar expression breadth.

We used a similar protocol to test which tissues show significant clustering. For each tissue, we calculated the dispersion of the gene density of all genes known to be expressed in this tissue. We then compared this with the corresponding dispersion values from 100,000 random permutations of the positions of all genes. To test whether any under-dispersion was caused by clustering according to expression breadth, we repeated these randomizations, this time permuting gene positions within each of the eight classes of similar breadth.

To estimate the range of co-expression clusters, we defined an index of co-expression ($ICE_{a,b}$) between two genes (a,b) as the number of tissues with common positive expression, weighted by the geometric mean of the two breadths (t runs over all tissues, $f_{a,t} \in \{0,1\}$ indicates not expressed/expressed):

$$ICE_{a,b} = \frac{\sum_t f_{a,t} f_{b,t}}{\sqrt{\left(\sum_t f_{a,t}\right)\left(\sum_t f_{b,t}\right)}}$$

Thus, $ICE_{a,b}$ ranges from 0 (no co-expression) to 1 (perfect co-expression). From this, we calculated a distance-based index of co-expression (ICE_d) as the mean of all gene pairs that are within a physical distance bracket $[d, d + 100\text{ kb}]$ apart on a chromosome. We compared ICE_d with the results expected under the null hypothesis (no spatial pattern in co-expression), by recalculating it for 100,000 data sets with randomly permuted gene positions.

To obtain size distributions of clusters, we defined a gene cluster as any contiguous group of co-expressed genes ($ICE_{a,b} \geq 0.5$ for all gene pairs a,b in the cluster). Cluster size was histogrammed according to cluster length (distance between the two outermost genes) in 10-kb windows. We compared this with random distributions from 1,000,000 permutations of gene positions.

URL. The UCSC Human Genome Project Working Draft, 6 August 2001 assembly (hg8) can be found at http://genome.cse.ucsc.edu/.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We acknowledge support by the Wellcome Trust (MJL), CONACyT and ORS (AOU), and BBSRC (LDH).

Competing interests statement

The authors declare that they have no competing financial interests.

Received 18 January; accepted 12 April 2002.

1. Ko, M.S.H. *et al.* Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. *Hum. Mol. Genet.* **7**, 1967–1978 (1998).
2. Bortoluzzi, S. *et al.* A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* **8**, 817–825 (1998).
3. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292 (2001).
4. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–488 (1995).
5. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
6. Blumenthal, T. Gene clusters and polycistronic transcription in eukaryotes. *BioEssays* **20**, 480–487 (1998).
7. The MHC Sequencing Consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
8. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Paul, A.L. & Ferl, R.J. Higher-order chromatin structure: looping long molecules. *Plant Mol. Biol.* **41**, 713–720 (1999).
10. Fisher, R.A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
11. Zavattari, P. *et al.* Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography chromosome recombination frequency and selection. *Hum. Mol. Genet.* **9**, 2947–2957 (2000).
12. Hurst, L.D., McVean, G.T. & Moore, T. Imprinted genes have few and small introns. *Nature Genet.* **12**, 234–237 (1996).
13. Schuler, G.D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
14. Lash, A.E. *et al.* SAGEmap: a public gene expression resource. *Genome Res.* **10**, 1051–1060 (2000).
15. Duret, L., Mouchiroud, D. & Gouy, M. HOVERGEN: a database of homologous vertebrate genes. *Nucleic. Acids. Res.* **22**, 2360–2365 (1994).

Chapter Five

Genome Structure and Gene Expression

Lercher, M. J., A. O. Urrutia, A. Pavlicek, and L. D. Hurst. 2003. A unification of mosaic structures in the human genome. Hum Mol Genet 12: 2411-5.

Human Molecular Genetics, 2003, Vol. 12, No. 19 **2411–2415**
DOI: 10.1093/hmg/ddg251

A unification of mosaic structures in the human genome

Martin J. Lercher[1,*], Araxi O. Urrutia[1], Adam Pavlíček[2] and Laurence D. Hurst[1]

[1]Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK and [2]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, 16637 Prague, Czech Republic

Received February 17, 2003; Revised and Accepted July 21, 2003

The human genome is a mosaic structure on many levels: there exist cytogenetic bands, GC composition bands (isochores) and clusters of broadly expressed genes. How might these inter-relate? It has been proposed that to optimize gene regulation, housekeeping genes should concentrate on transcriptionally competent chromosomal domains. Prior evidence suggests that regions of high GC and R bands are associated with such domains. Here we report that broadly expressed genes cluster in regions of high GC, and in R and lightest Giemsa bands. This is not only a confirmation of the adaptive hypothesis, but is also the first direct systematic evidence of a general interdependence of expression patterns with base composition and chromosome structure.

INTRODUCTION

What determines gene order in the human genome? Genes are not randomly distributed along chromosomes. We have recently shown that they are arranged according to their breadth of expression: broadly expressed genes tend to cluster (1), although factors that account for this clustering remain unknown. In prokaryotes, genes related to a particular function are clustered in operon structures and their expression is co-regulated. While in eukaryotes co-regulatory gene units have been observed, in some cases, as in the case of HOX genes, there is no evidence for these to be a common case.

Unlike prokaryotes and other invertebrates, mammalian genomes show great variability in their base composition (2). Several hypotheses have been proposed to explain this pattern. Some authors favouring a selectionist explanation argue that high contents of G + C in some regions of the genome help to preserve chromatin structure in thermo-regulated organisms (2). Theories of mutational processes to explain base compositional differences have also been proposed (3). Nevertheless, the reason for the heterogeneity in base composition is still a matter of debate. How, if at all, do the compositional mosaic structure of the genome and the gene expression patterns interact?

If selectively neutral processes determine both the mosaic structure of chromosomes and the clustering of broadly expressed genes, then we expect no relationship between regional composition and functional properties of the genes such as their expression patterns. On the other hand, regions differing in their base composition may be differently suitable for transcription. If local chromatin characteristics affect access to the transcription machinery (4–6), then we expect genes expressed in many cell types to be concentrated in transcriptionally competent regions, even when gene density effects are corrected for.

It is well known that chromosomal regions of high GC exhibit higher gene densities (7). These regions also contain a higher density of CpG islands (8). Because it has been reported that housekeeping genes—in contrast to tissue-specific genes—are always associated with CpG islands (9), this has led to the widely accepted notion that housekeeping genes are preferentially located in regions of high GC (2). However, a detailed analysis found that the association between CpG islands and the expression patterns of genes is more complex: 10% of housekeeping genes are not associated with CpG islands, while this fraction varies for tissue-specific genes between GC-poor and GC-rich regions (10). Furthermore, the latter study concluded that housekeeping genes are slightly more prevalent in GC poor regions, once gene density has been accounted for. Thus, this systematic study (as well as two others from the same group) (11,12) contradicts widely held beliefs on the association between expression breadth and regional nucleotide composition. However, these reports measured expression breadth from expressed sequence tag (EST) data, and GC content from coding sequences; both are not ideal measures. Thus, the question of how housekeeping genes are distributed in relation to tissue specific genes in the human genome is currently not fully resolved.

*To whom correspondence should be addressed. Tel: +44 1225385902; Fax: +44 1225386779; Email: m.j.lercher@bath.ac.uk

RESULTS

Our aim is to evaluate whether such a relation between regional base composition and gene expression exists. Until recently it was not possible to systematically address this question due to the lack of reliable quantitative expression data necessary to discriminate expression rate from expression breadth. Serial Analysis of Gene Expression (SAGE) technology (13) allows quantitative identification of genes expressed in a particular tissue. To examine whether gene order in the genome is related to base composition variation, we compared expression patterns of over 10 000 autosomal human genes across 19 normal tissues with the GC content of their introns. It has recently been shown that under some experimental conditions, SAGE libraries may tend to over-represent GC rich sequences (14). As this could bias our results, all analyses are based on a curated dataset, which excludes libraries that showed a bias towards GC rich sequences (see Materials and Methods).

There appear to be two types of models that predict a correlation between local chromatin characteristics and expression pattern. The first type assumes that chromatin remodelling acts like a switch, either allowing or preventing the transcription of genes. This would predict a correlation of GC and banding pattern with expression breadth (the number of tissues where a gene is expressed), but not with measures of expression rate. The second type of model assumes that chromatin remodelling dominantly affects the rate of transcription, e.g. by ensuring that highly expressed genes (be they tissue-specific or broadly expressed) are in open chromatin. This model would predict an association of chromatin characteristics with peak expression rate, but not necessarily with expression breadth. To distinguish between these two models, we report results for both of these measures (1): breadth of expression and peak rate of expression. We also performed corresponding analyses for other measures of expression rate (mean across all tissues, mean across tissues with positive expression, standard deviation over mean across all tissues), although we are not aware of a model that would predict a direct effect on these measures. All measures of expression rate are highly correlated, and all results are in qualitative agreement with those presented here for the peak rate (data not shown).

Analysis of expression breadth and local nucleotide composition (GC) reveals a highly significant correlation ($r^2 = 0.24$, $P < 10^{-5}$; for an example see Fig. 1). A similar although weaker pattern appears when comparing GC content and the logarithm of the expression rate ($r^2 = 0.05$, $P < 10^{-5}$). To account for the great degree of variability in expression patterns at a one-gene resolution, these correlations were assessed after averaging all variables over 15 neighbouring genes. Furthermore, after sorting individual genes according to their surrounding DNA composition into GC categories of 5% width, mean expression breadth and log(rate) both have a strikingly strong linear relationship with base composition ($r^2 = 0.89$, $P < 0.0005$; $r^2 = 0.83$, $P < 0.005$, respectively; Fig. 2). We previously reported a limited although significant correlation between expression patterns and base composition on a one-gene basis (1). Correlation coefficients rise as the number of genes per window is increased (Fig. 3; all correlations are highly significant, $P < 0.0005$). Thus, while much of the variation in expression breadth and rate is based in the properties of individual genes, a large fraction of the *long-scale* variability (up to almost 50%, see Fig. 3) is predicted by a related variation in GC composition. This strongly supports the notion that isochores are real and may have some functional importance.

Our earlier analyses showed that clustering of genes was related to expression breadth and that the previously described clustering of highly expressed genes (15) is a by-product of the dependence of rate on breadth (1). Accordingly we found that the correlation of log(rate) with GC content fades out when we look at residuals from the breadth correlation. In contrast, when examining the residuals from breadth on log(rate), the correlation with GC remains unchanged (Table 1). These results provide evidence for a strong relationship between breadth of expression of a gene and the base composition at the genomic region where it is situated.

In contrast to the above results, some previous analyses have reported a small *negative* correlation between local GC content and the breadth of expression estimated from expressed sequence tag (EST) data (10–12). To reconfirm that our results are not an artefact of the SAGE method, we therefore repeated our analysis using the breadth of expression obtained from the ESTs contained in the UniGene database (16). In qualitative agreement with the SAGE analysis in Figure 3, we found a highly significant positive correlation between intron GC and EST breadth of expression, which increased with the number of neighbouring genes averaged (Supplementary Material Figure A). The discrepancy between our results and previous studies appears to be caused mainly by the previous studies examining individual genes rather than regional averages. Another contribution to this difference may stem from the use of (total or third site) coding sequence GC instead of intron GC; coding region and intron GC appear to measure different genomic properties. However, we found qualitatively similar results for intergenic GC, intron GC excluding repetitive sequence and transcript GC (data not shown). It has been suggested that a discrepancy between SAGE and EST results might be due to a differential decay of SAGE tags with different GC (12). This appears not to be relevant: there is hardly any correlation between SAGE tag GC and expression breadth in our curated data set ($r^2 = 0.0001$).

GC content has been associated with CpG density. Given that housekeeping genes tend to be located near CpG islands (10,17), the concentration of housekeeping genes was expected to be higher in GC rich regions (2). This suggests a possible explanation for our findings, i.e. the correlation between expression breadth and GC content might simply reflect the higher CpG density rather than GC content *per se*. However, we found very similar results when correlating expression breadth with intron GC excluding CpG islands ($r^2 = 0.79$ for 5% bins of GC). Thus, CpG island preference alone fails to explain the concentration of housekeeping genes in GC rich regions. From the above we might presume that isochores are, to a very large extent, regions of comparable breadth of expression.

The mammalian genome is also heterogeneous in its structure. Giemsa staining of metaphasic chromosomes reveals a banding pattern. The Giemsa bands are related to chromatin compaction and distribution of chromosomes inside the

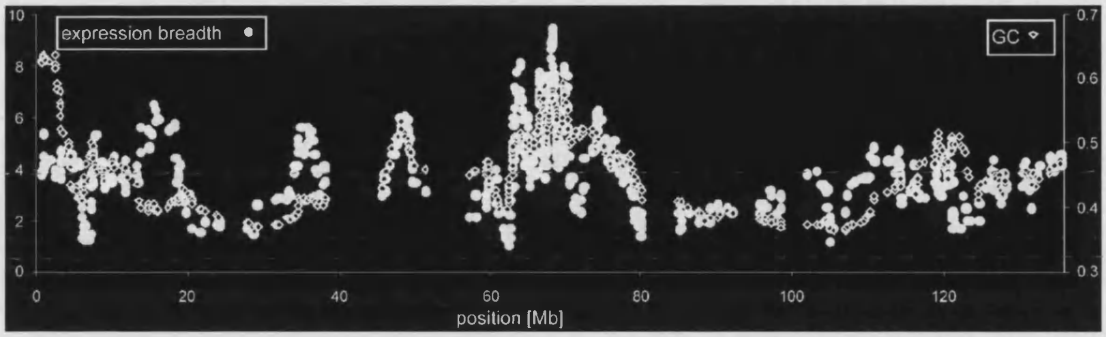


Figure 1. Expression breadth (black dots) and intron GC (grey diamonds) for genes on chromosome 11. Each point represents the average of GC content /breadth for 15 neighbouring genes.

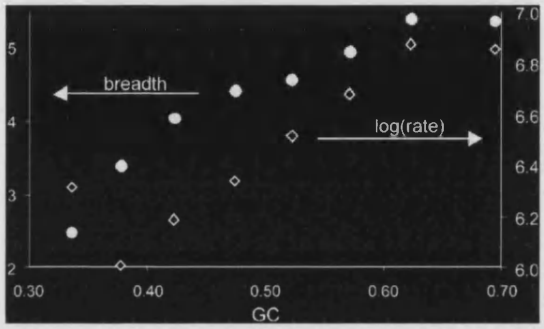


Figure 2. Expression breadth and log(rate) averaged over contiguous intron GC windows of 5% width. The correlation coefficients give $r^2 = 0.89$ (breadth) and 0.83 (rate), respectively.

Table 1. Correlations for rate and breadth with base composition; 15-gene averages

	r	r^2	P
Breadth versus log(rate)	0.43	0.19	$<10^{-5}$
Breadth versus GC	0.49	0.24	$<10^{-5}$
Residuals of breadth = $a + b \times$ log(rate) versus GC	0.43	0.19	$<10^{-5}$
Log(rate) versus GC	0.23	0.05	$<10^{-5}$
Residuals of log(rate) = $a + b \times$ breadth versus GC	0.02	0.0005	0.58

nucleus, where darker and more compacted regions tend to occupy the nuclear periphery (4). Moreover, band types have been correlated with base composition: GC-poorest DNA segments are preferentially located on the most intensely staining G bands, while a subset of the R bands contains the GC-richest isochores (18,19). Therefore we asked whether clustering of housekeeping genes in GC-rich regions relates to these chromosome bands. Indeed, we found that broadly expressed genes are preferentially located in the lightest staining G and R bands (Fig. 4), which contain the most GC-rich segments. Overall 81% of housekeeping genes (expressed in 13 or more tissues) are in one of these two bands. Gene density is generally higher in these two bands (19,20); nonetheless, controlling for gene density we still find enrichment of broadly expressed genes in the R and lightest staining G bands (747 genes compared with 687 expected; $P = 0.023$ from χ^2 test).

The observed mean expression breadth decreases much steeper from R- to dark G-bands than predictions derived from either total band GC or from the intron GC of the genes under study (Fig. 4). This suggests that at least part of the correlation between banding patterns and expression breadth is independent of GC. Consequently, examining the regression residuals of expression breadth versus intron GC for individual genes, we find that genes are not randomly distributed across cytogenetic bands (ANOVA; $P = 0.038$ from F-test). Thus, broadly expressed genes show independent preferences for regions of high GC as well as for the R and lightest staining G bands.

DISCUSSION

Our results provide the first direct systematic evidence of a general relationship between expression patterns and chromatin structures and base composition. This however leaves unresolved the issue of the evolution of isochores. Might GC content evolve as a by-product? Or is it necessary that regions of broad expression have a high GC content, i.e. is the GC content itself under selection? Assuming that housekeeping genes tend to concentrate in regions of open chromatin in order to facilitate transcription (4), our data could be consistent with two models that explain the higher GC content in DNA segments containing housekeeping genes. In the first model, GC content is selectively driven since GC-rich DNA tends to be open and taken to the centre of the nucleus. Alternatively, high GC content could, via biased gene conversion, be a by-product of open chromatin being more prone to recombination.

Both models are consistent with the correlation between recombination rates and base composition (21–25). In the former model this would be a side consequence of the fact that open chromatin is GC rich and open chromatin may be prone to recombination. In the latter model, the GC content is

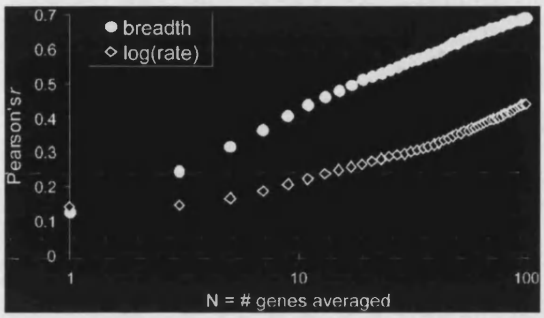


Figure 3. Pearson's r for the correlation between intron GC and expression breadth and rate, for sliding averages over N neighbouring genes.

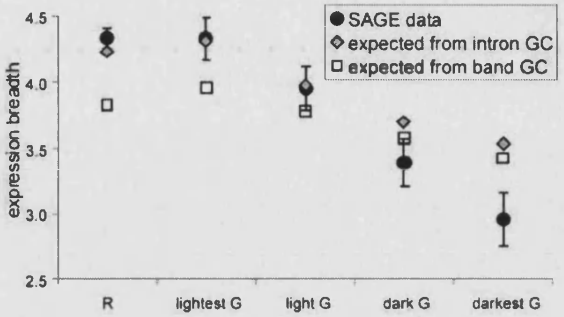


Figure 4. Mean expression breadth of genes in differently staining cytogenetic bands and predictions from intron GC and from total band GC. Error bars show standard errors of the means.

caused by recombination. Therefore both models are also consistent with a correlation between breadth and recombination. Indeed we find such pattern, although the correlation is extremely weak, possibly due to the low resolution of the data available ($r^2 = 0.0034$, $P < 10^{-4}$ for 15-gene averages; recombination data from 25). However, we can imagine a discriminating prediction. Under the second model, all genes expressed exclusively in germ cells just prior to chiasmata formation are prone to recombination and hence to high GC content, while the former predicts that, as such genes are tissue specific, they need not be GC rich. When SAGE libraries for these cell types become available, the test could be performed.

How might the association between expression patterns and local chromatin characteristics shown above be tested experimentally? The above model predicts that when genes are inserted into a non-native chromosomal environment together with their promoter regions, their expression pattern should depend on local GC content and cytogenetic banding pattern. It is indeed well known that randomly inserted transgenes are often not transcribed. In agreement with the competent chromatin model, transgene expression—at least in the case of globin genes—can be rescued with locus control region elements that modify chromatin structure (26). By a systematic examination of the local chromatin characteristics and the expression pattern for a large number of randomly located transgene insertions, the predictions of our model can thus in principle be tested. Unfortunately, currently available data is not of adequately high resolution to address this issue (F. Grosveld, personal communication), and we have to leave this test for future work.

In summary, our results are consistent with gene location being an adaptive property related to regional base composition and chromosome structure (2), where selective pressures favour the concentration of housekeeping genes in genomic regions with particular structural properties, most probably to facilitate access to transcription machinery (4). In accord with this picture, it has been shown that actively transcribed chromatin is predominantly located within the nuclear interior comprising early replicating R bands, which contain the GC richest and gene richest domains (27). The null model, in which genes in the genome are randomly assorted with respect to their expression, is no longer tenable.

MATERIALS AND METHODS

The Serial Analysis of Gene Expression (13) (SAGE) data was obtained from SAGEmap (28) (ftp://ncbi.nlm.nih.gov/pub/sage). The dataset was curated to avoid possible GC biases in SAGE libraries following the approach of Margulies *et al.* (14); we removed 14 libraries with mean tag GC > 0.5. The resulting SAGE tag/tissue data set was based on 40 libraries representing 19 tissues. Tag counts were converted to relative values (cpm, counts per million) after joining all libraries representing the same tissue type. If tags were found only once in one tissue type, we discarded the observation as a likely sequencing error. This data was cross-linked to the mRNA sequences in RefSeq (ftp://ncbi.nlm.nih.gov/refseq), by extracting the 3′-most *Nla*III SAGE tag for each mRNA. If the same tag occurred more than once in RefSeq, all corresponding genes were excluded. To be conservative, the gene set was further restricted to those sequences who's tag was also reported by NCBI as reliable for the corresponding UniGene cluster (16) (UniGene build #155, ftp://ncbi.nih.gov/pub/sage/map/Hs/NlIII/SAGEmap_tag_ug-rel.zip). For the remaining genes, we calculated breadth of expression as the number of tissues with positive expression. For genes expressed in at least one tissue, we also calculated the peak rate of expression (maximum cpm across tissues). As with all forms of expression assay, the SAGE data employed here will inevitably miss some genes expressed at low levels. However, this is not likely to unduly bias our results: as we have demonstrated earlier (1), controlling for rate of expression hardly affects regional variation in expression breadth.

Of the genes with valid expression information, 10774 could be located unambiguously on the June 2002 UCSC genome assembly (29) (ftp://genome-archive.cse.ucsc.edu). Gene position was defined as the midpoint between 5′ and 3′ ends of the transcribed sequence.

For each gene, we extracted the coding sequence from the RefSeq mRNA. We also extracted transcripts (containing both exon and intron sequences, and including information on repetitive DNA) from the genomic data at the UCSC web site. Owing to sequencing errors, mistakes in the assembly, or mis-annotations, intron sequences may be wrongly identified from this kind of data. To ensure proper identification, we compared the coding part of the corresponding exons against the RefSeq sequences. Genes were excluded if we found a length difference or if an internal stop codon occurred in the genomic

coding sequence. Nucleotide composition was measured as the guanine and cytosine (GC) fraction. Intron GC was calculated for 8128 genes with total intron length >100 bp. For 7986 genes with total intron length >500 bp, we also calculated intron GC excluding CpG islands. CpG islands were defined as regions of at least 200 bp, with mean GC > 0.5, and CpG observed/CpG expected >0.6 (10).

Recombination data (25) and cytogenetic band positions (based on FISH data) (30) were also obtained from the UCSC web site. Band positions are imprecise by up to several 100 kb or even more. When including only genes at least 1 Mb away from start and end of their cytogenetic band, results are qualitatively unchanged (data not shown).

To reconfirm that the observed patterns are not due to any remaining bias of the SAGE data, we also examined the correlation between nucleotide composition and local breadth of expression obtained from expressed sequence tag (EST) data. Each UniGene group not only contains the RefSeq mRNA sequence, but also all ESTs believed to map to the same gene. We used these to cross-link genes to 622 EST libraries constructed from normal tissue samples, each containing at least 50 ESTs. This resulted in a data set of 8763 genes, each known to be expressed in at least one out of 73 normal tissues (16 prenatal and 57 postnatal). We calculated breadth of expression as the number of tissues with positive expression information.

For all correlations, r is Pearson's coefficient. Significance levels were estimated from 10 000 random pairings of the raw data value pairs: $P = (1 + \text{number of random pairings with smaller or equal } r^2)/(1 + \text{number of random pairings})$. Correlations and regressions for expression rate were calculated after taking the logarithm.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Laurent Duret and Frank Grosveld for interesting discussions. This work was supported by CONACyT and ORS (A.O.U.), BBSRC (L.D.H.), and The Wellcome Trust (M.J.L.).

REFERENCES

1. Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.*, **31**, 180–183.
2. Bernardi, G. (1993) The vertebrate genome: isochores and evolution. *Mol. Biol. Evol.*, **10**, 186–204.
3. Gu, X. and Li, W.H. (1994) A model for the correlation of mutation-rate with GC content and the origin of GC-rich isochores. *J. Mol. Evol.*, **38**, 468–475.
4. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
5. Mahy, N.L., Perry, P.E. and Bickmore, W.A. (2002) Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *J. Cell Biol.*, **159**, 753–763.
6. Williams, R.R. (2003) Transcription and the territory: the ins and outs of gene positioning. *Trends Genet.*, **19**, 298–302.
7. IHGSC (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
8. Cross, S.H., Clark, V.H., Simmen, M.W., Bickmore, W.A., Maroon, H., Langford, C.F., Carter, N.P. and Bird, A.P. (2000) CpG island libraries from human Chromosomes 18 and 22: landmarks for novel genes. *Mamm. Gen.*, **11**, 373–383.
9. Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
10. Ponger, L., Duret, L. and Mouchiroud, D. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.*, **11**, 1854–1860.
11. Goncalves, I., Duret, L. and Mouchiroud, D. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.*, **10**, 672–678.
12. Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–649.
13. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–488.
14. Margulies, E.H., Kardia, S.L. and Innis, J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucl. Acids Res.*, **29**, e60.
15. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A. *et al.* (2001) The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
16. Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540–546.
17. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
18. Saccone, S., Desario, A., Wiegant, J., Raap, A.K., Dellavalle, G. and Bernardi, G. (1993) Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA*, **90**, 11929–11933.
19. Federico, C., Andreozzi, L., Saccone, S. and Bernardi, G. (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Res.*, **8**, 737–746.
20. Saccone, S., Federico, C., Solovei, I., Croquette, M.F., Della Valle, G. and Bernardi, G. (1999) Identification of the gene-richest bands in human prometaphase chromosomes. *Chromosome Res.*, **7**, 379–386.
21. Bernardi, G. (1989) The Isochore organization of the human genome. *A. Rev. Genet.*, **23**, 637–661.
22. Ikemura, T. and Wada, K.-N. (1991) Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromsome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucl. Acids Res.*, **16**, 4333–4339.
23. Holmquist, G.P. (1992) Chromosome bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.*, **51**, 17–37.
24. Fullerton, S.M., Carvalho, A.B. and Clark, A.G. (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.*, **18**, 1139–1142.
25. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
26. Grosveld, F., van Assendelft, G.B., Greaves, D.R. and Kollias, G. (1987) Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*, **51**, 975–985.
27. Sadoni, N., Langer, S., Fauth, C., Bernardi, G., Cremer, T., Turner, B.M. and Zink, D. (1999) Nuclear organization of mammalian genomes. Polar chromosome territories build up functionally distinct higher order compartments. *J. Cell Biol.*, **146**, 1211–1226.
28. Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J. and Altschul, S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
29. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
30. Cheung, V.G., Nowak, N., Jang, W., Kirsch, I.R., Zhao, S., Chen, X.N., Furey, T.S., Kim, U.J., Kuo, W.L., Olivier, M. *et al.* (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, **409**, 953–958.

Chapter Six

General discussion

DISCUSSION

In this thesis I have addressed the possible role of selective pressures associated with protein synthesis in the evolution of human genes. At the time the work presented here begun, the most widely accepted view was that mammalian genes were not significantly affected by pressures related to gene transcription or translation efficiency; this was because of the relatively small sizes of mammal populations. Using bioinformatics tools on large samples of human genes, I have shown gene expression to be, unexpectedly, related to a variety of gene sequence parameters. Highly expressed genes are encoded by shorter genes and tend to have shorter intervening sequences. Codon and amino acid usage show a higher deviation from random expectations in genes that are highly transcribed. All of these patterns are difficult to explain under a neutral scenario and are consistent with expression mediated selection on gene sequences. In addition, evidence for a role of chromosome organisation and structure in gene expression regulation has been observed. Highly expressed genes are clustered in the genome and tend to concentrate in regions of high gene density. These gene distribution patterns are closely related to structural characteristics of the genome such as base composition and Giemsa bands. Highly expressed genes tend to concentrate in regions of higher G+C content and lightest Giemsa bands that characterise open chromatin. These are presumed to be located towards the centre of the nucleus. These observations are in accordance with a previously proposed hypothesis of transcription competency heterogeneity along chromosomes, and subsequent selection on genes (particularly on highly expressed genes) to concentrate in these regions.

Codon usage has been associated with expression levels in non-mammalian species where highly expressed genes tend to have higher biases in the use of alternative codons (Duret and Mouchiroud 1999; Gouy and Gautier 1982; Sharp et al. 1986; Stenico et al. 1994). These results are consistent with selection acting on silent sites to optimise protein synthesis. I analysed this parameter in a sample of human genes in Chapters 2 and 3. Chapter 2 presented a novel index for measuring codon bias (MCB) that corrects for background base composition. So how effective it is for measuring codon bias? In Chapter 2 MCB is shown to correct for background variation in nucleotide composition, and additional analysis showed that it is highly insensitive to variations in amino acid composition and rare amino acids. MCB index has been independently reviewed and compared against other methods used to calculate codon bias (Novembre 2002). The results of this analysis confirmed that MCB correctly accounts for background base composition. As for the majority of codon bias indexes, however, the method is unfortunately strongly dependent on gene length. Therefore this variable should be corrected for when interpreting MCB indexes. Because the index can be used to correct for base composition variation, it may be suitable for interspecies comparisons.

From the publication of the method of MCB to the time of this writing, more indexes for measuring codon bias have been added to the list (see for example Novembre 2002). From personal communications I am aware of even more being developed. This diversity of indexes partly stems from the difficulties of accurately measuring codon bias.

So, what exactly are we trying to measure? The use of alternative codons can be thought of as an ensemble of several random variables: one per amino acid. Each of these variables has 2-6 possible different outcomes or codons (amino acids encoded by only one codon cannot have codon usage bias). Each outcome (or codon) has an associated

probability of appearance within the sequence. For any given gene, biases for all 18 amino acids (Methionine and Tryptophan are encoded by only one codon) can be arranged in a vector (or observed distribution) which can in turn be compared to the expected distribution. We can then obtain an index of departure from expectations for each amino acid in a fairly straightforward manner. In order to obtain a single-value index of codon usage bias for a particular gene, however, biases of individual amino acids have to be added in a sensible way. And it is this step that is the more problematic one. Different amino acids within a gene vary in two aspects: number of times a particular amino acid appears within the sequence and their degree of degeneracy. If an amino acid is rare, then the observed distribution is more likely to be far from the expected just by chance, therefore the bias of a rare amino acid should have less impact on the overall index of codon bias. In addition, it requires a greater selective pressure to cause a particular codon to take 90% of the times of appearance of the amino acid it encodes, for a four fold than a two fold degenerate amino acid. Therefore, biases in four fold degenerate amino acids should have a higher weight.

Is a perfect codon bias estimate likely to be developed? Given all these considerations, it is my thought that a near-perfect method for estimating codon bias as departure of randomness of codon distributions, which can cope with differences in gene length, variation in amino acid proportions and base composition, might not be constructed in the near future. As computer power increases, then strict maximum likelihood indexes might provide a more accurate codon bias estimate.

The initial analysis of the relationship of codon bias presented in Chapter 2 showed that over 80% of the genes do have a higher codon bias than that expected from their base composition. In addition, after correcting for background nucleotide composition, most genes tend to favour the same set of codons. However I found only a relatively weak

correlation between codon bias and gene expression. This was mostly explained by gene length variation. Setting aside the considerations on measuring codon bias mentioned above, the lack of evidence for codon bias as being influenced by expression patterns could derive from the type of data used to estimate expression. Here, breadth of expression was used as an estimate of expression levels. The number of tissues where a gene is being expressed might not necessarily be a good estimate of its level of expression.

In Chapter Three a re-evaluation of codon bias is presented using a larger sample size and quantitative gene expression data. The relationship between codon bias and expression is maintained even after correction for gene length. Is codon bias related to expression patterns in human genes? Biases in mutational patterns favouring specific nucleotides over others may influence codon choice. Those processes coupled with transcription frequency, in particular might explain the relationship I found between codon bias and expression patterns. GC content is higher in highly expressed genes (Urrutia and Hurst 2003). This expression dependence of base composition could explain a possible relationship between expression and the use of alternative codons. In addition, Green et al. (2003) showed that there is a mutation asymmetry in the DNA transcribed strand compared to the non-transcribed strand, favouring nucleotide changes to G and T. A further analysis performed on a sample of human genes (Majewski 2003) showed that nucleotide composition is in fact related to expression patterns. However, because the MCB method corrects for background nucleotide compositional biases, these do not increase the codon bias index. Therefore, my results cannot be explained by nucleotide biases even when related to expression patterns of genes.

In comparison to other species, such as *S. cerevisiae*, the strength of the relationship between expression and codon bias is relatively moderate in human genes. One possible

reason for this difference is that the use of coding regions to estimate expected codon distributions might be too conservative. If similarly ending codons are favoured (as is the case of Drosophila), then expected distributions based on coding regions will be closer to the actual distribution and codon bias would be underestimated. The use of intergenic regions, on the other hand, might produce an overestimation of codon usage bias. This is because transcription-coupled mutation/repair processes may cause a departure of base composition of transcribed sequences from surrounding intergenic regions. Using intronic regions to obtain expected codon distributions might provide a better estimate. Intronic regions are subject to the same transcription rates as coding regions but are not translated. Therefore these sequences should be more suitable for setting expected distributions to examine translational selection (although codon bias might partly reflect selection acting over transcription efficiency as well Vinogradov 2001a; Vinogradov 2003). Caution should be paid to the impact of transposable elements and functional elements within introns that compose an important percentage of intronic sequences.

In Chapter Three, I evaluated the relationship between expression patterns and other sequence characteristics. A recent large scale study in human and *C. elegans* genes (Castillo-Davis et al. 2002) had shown that intron content is related to expression levels i.e. highly expressed genes have short introns. This result could be interpreted as the action of selection to reduce the amount of sequence to be transcribed. This result suggests that selection might be acting to reduce the cost of transcribing long introns in highly expressed genes in mammalian genes. However, when I examined the relation between length of intron and intergenic regions, I found that intron length is strongly related to intergenic distances. Therefore, the Castillo-Davis (2002) study does not address whether the observed pattern is the result of a direct link between expression levels and intron length or

a by-product of a general compaction in regions where highly expressed genes reside. To evaluate whether the reduced intron sizes found in highly expressed genes were compatible with transcription cost reduction, rather than with a regional compaction, I repeated the analysis after correcting for regional effects. I found that intron sequences are significantly influenced by expression patterns even after correcting for intergenic distance. Length of coding region is also strongly influenced by expression patterns and, as expected, the relation between expression levels and protein size is little affected by regional effects.

The above results suggest that both regional genomic parameters as well as expression mediated pressures determine resulting intron sizes and coding regions. The relation between intergenic distances with gene characteristics shows that regional mutation patterns and indel ratios are a significant factor shaping gene sequences and should be taken into account when evaluating the input of other variables.

Highly expressed genes benefit from lower intron content. That is, the total number of bp constituted by introns in each gene is lower in highly expressed genes. But how do genes achieve this reduction? One possibility is a reduction in the number of introns. Consistent with this, highly expressed genes have fewer introns. However, in fact, number of introns per bp is not strongly related to expression levels (Castillo-Davis et al. 2002). Therefore the reduction in intron number in highly expressed genes is a by-product of the reduction in coding sequence size of highly expressed genes. The observed reduction in intron content in highly expressed genes is the result of the reduced size of individual introns. This leaves us to explain the prevalence of introns in mammalian genes. This issue goes beyond the scope of this thesis but there are a number of possible reasons for the maintenance of introns. Introns might facilitate alternative splicing which is very common in mammalian genes (Hickey and Benkel 1986), increase recombination between exons

(Comeron and Kreitman 2000) and/or correct chromatin structure to facilitate transcription (Vinogradov 2001a; Vinogradov 2003). Interestingly, Lynch and Kewalramani (2003) recently proposed that introns, regularly spaced along genes, might facilitate the identification of aberrant RNA sequences with early termination codons to be degraded before translation; this operates by the coupling of an RNA surveillance mechanism that requires nearby exon junctions to recognize premature termination codons.

Interestingly, I also found that amino acid composition is significantly influenced by expression levels. The frequency of the majority of amino acids is correlated with expression level. Dufton (1997) calculated amino acid complexity indexes in terms of molecular weight and tri-dimentional conformation and found that low complexity amino acids are more common in proteins. Akashi and Gojobori (2002) obtained similar results by calculating metabolic costs of producing each amino acid in the bateria *E. coli* and *B. subtilis*. In addition, they found that amino acids of lower metabolic cost tend to be more common in genes with high codon bias. In accordance with a selection scenario for cheap amino acids, I found that human highly expressed genes tend to encode a larger proportion of low complexity amino acids (as defined by Dufton 1997). More accurate estimates of the true metabolic cost of production/acquisition of amino acids in mammals are needed to reveal the extent of these pressures in shaping amino acid choice.

The studies presented in Chapters Two and Three are among the first studies that systematically address the relationship between expression patterns and gene characteristics, with the aim of evaluating whether human genes show significant signatures of selection related to protein synthesis efficiency. The results obtained support the conclusion that, despite previous expectations, human genes possess signatures compatible

with an optimization of protein synthesis related costs both at the transcriptional and translational level.

The release of the human genome sequence allowed further study of the effects of expression patterns on genes even beyond the gene sequences themselves. Chapter Four presents the analysis of gene sorting with respect to expression patterns. The results show that genes are sorted according to their expression. Broadly expressed genes are more likely to have neighbours which are also broadly expressed rather than lowly expressed ones. I examined further gene distribution along chromosomes in order to investigate the reasons for the non-uniformity of gene distribution with respect to their expression patterns. In particular, I analysed the possibility of the clustering of broadly expressed genes to be the result of heterogeneity in transcription competency along chromosomes as proposed by Cremer & Cremer (2001). If different regions of the genome have different transcription competencies, then these differences should affect all genes, since they also need to be expressed. The amount of selective pressure to be located in regions of higher transcription competency, however, would vary from gene to gene according to their breadth of activity. From this we can draw two main expectations; first, we should observe heterogeneity in gene densities along the chromosome. Second, since broadly expressed genes are under greater pressure to be located in regions of higher transcription competency, they should tend to be over represented in regions of higher gene density.

It is well known that gene density is highly heterogeneous along chromosomes (Mouchiroud et al. 1991); however, gene density heterogeneity cannot in itself be proof of transcription related selection. Neutral processes related to variation in indel ratios could result in such a distribution. If gene densities along chromosomes are determined by neutral processes, unrelated to expression regulation, then we should find highly expressed genes

to be equally distributed among high and low gene density regions. Alternatively, for example, if genes in densely packed areas tend to interfere with one another by competing for transcription factors then we would expect to find highly expressed genes preferentially located in lowly populated areas of chromosomes. One possible way to test this is, for example, to take a subset ten percent of genes with highest expression rates, then one in every ten genes located in regions of low density should correspond to our subset, and the same when we look in regions of high gene density. If on the other hand, gene densities are related to expression patterns of genes, then the above distributions should not be recovered. Chapter Three shows that highly expressed genes in regions of high gene density are over represented relative to random expectations. These observations are consistent with the notion that some chromosomal regions are more suitable for transcription. Also it suggests that gene density might be at least partly determined by selection for occupying more suitable regions.

The above analyses suggest that genes are not distributed randomly along the genome but that their location is related to their levels of activity possibly to favour their location in regions of higher transcription competency. However, these analyses do not address what actually makes a region more suitable for transcription. The human genome is composed of 23 pairs of chromosomes all densely packed inside the nucleus, and chromosome regions vary in their nuclear location. Cremer and Cremer (2001) have proposed that genes in chromosome regions located in more interior parts of the nucleus have better access to the transcription machinery than genes located in regions in the periphery of the nucleus. This being so, genes with broader expression would be expected to concentrate in competent regions. Performing an experimental assessment of transcription rates of genes located in different parts within the nucleus is technically

difficult (particularly for an *in silico* graduate student). An alternative approach to test Cremer & Cremer's hypothesis is to examine the relation between expression patterns with variables related to chromosome structure. Chromosome Giemsa staining reveals a striped pattern of chromosome banding. These bands have been related to chromosome nuclear arrangements (Ferreira et al. 1997). Darker bands of highly compacted chromatin tend to be located at the periphery. In addition, it has been observed that lighter Giemsa bands of open chromatin possess a higher G+C content (Saccone et al. 1993). In Chapter Five I analysed in detail the relationship between expression levels of genes with base composition and chromosome banding patterns.

I show that expression patterns of genes are strongly related to regional non-coding base composition of both intron and intergenic regions: highly expressed genes concentrate in regions of high G+C content. Furthermore, Giemsa bands are also negatively related to expression patterns of genes. These results suggest that, as Cremer & Cremer (2001) have proposed, broadly expressed genes are preferentially located in open chromatin closer to the centre of nucleus.

Together the results presented in this thesis show that, contrary to expectations, genes show clear signs of significant selective pressure for the optimization of protein synthesis, i.e. broadly/highly expressed genes tend to encode for shorter proteins, possess smaller intron sequences, have higher codon bias and are biased towards the encoding of cheaper amino acids. In addition, evidence supporting a relation between gene location and genome structural features is presented. Highly expressed genes are located in highly gene dense areas of higher G+C content open chromatin, presumably at the centre of the nucleus. These observations show that chromosome structure may play a more important role in gene regulation than previously thought.

The validity of the results here presented depends on the accuracy of expression data estimates. Errors in estimating expression levels of individual genes are likely to increase the noise of the data and obscure the patterns found. In this thesis, different sources of gene expression data have been used. These include EST, SAGE and chip-array data. All of these datasets present different technical limitations and types of errors which have been discussed in the Introduction. I compared expression data for the two datasets that allow estimation of expression levels, SAGE and chip array. For 10 tissues, there are data available using the two datasets (whole brain, lung, liver, pancreas, ovary, prostate, spinal cord, cerebellum, heart, kidney). I found that expression indexes are correlated by an r=0.344 on average (but see Ishii et al. 2000) for a direct experimental comparison). Differences may partly derive from inaccurate assignment of sequence tag for SAGE or unspecific oligo binding to mRNAs for chip array data. An additional source of variability between samples corresponding to the same tissue comes from the fact that tissue samples from individuals with different death cause, age and gender are used as well as from variations in the particular sections of tissue extracted. Indeed, when comparing expression data for different libraries using SAGE methodology I found that gene expression level estimates differ (r=.724 on average). I expect that, with the availability of more accurate expression, estimates the signals here recovered would become stronger.

Some forms of systematic bias in the acquisition of expression estimates, however, have been identified and could in fact contribute to the patterns observed (Balazsi et al. 2003). As discussed in the Introduction, these sources of error are specific to the method used to recover expression data. It is for this reason that we have refrained from any attempt to use composite estimates when data were obtained using different techniques. In the manuscripts presented in this thesis, at least two independent datasets of gene

expression have been used. This approach increases the validity of observations using expression estimates (Detours et al. 2003). A recent study has used an integrative approach for obtaining expression estimates from SAGE and chip array and compared them with gene length and base composition, reaching similar conclusions (Versteeg et al. 2003).

The aim of the work I have reported was to find out whether human genes showed signs consistent with the action of selection related to protein synthesis cost optimization. The results therefore add to the understanding of gene sequence evolution and the impact of genome structure on gene regulation in human genes. However they can also be compared to other species along the evolutionary scale (by this, I do not try to imply progression), to get a better perspective of the evolution of the forces shaping genes and genomes.

Codon usage bias, for example, has been strongly related to expression patterns in yeast. In this species, codon bias is a very good predictor of expression levels of genes. In other non-vertebrate species, codon bias is also strongly related to levels of expression. Direct measures of expression levels in bacterial species for large number of genes are not available, but there is no evidence of a different picture to that found from non-vertebrate eukaryotes. In contrast, in vertebrates -at least in mammals- the relationship between codon bias and expression levels is significantly weaker. Here, regional base composition appears to be a major determinant of synonymous sites. In fact, even amino acid determining codon positions also tend to match regional base composition in mammalian genes but to a lesser extent (Clay et al. 1996).

Codon bias also appears to shift in terms of the main factors determining it in different species. In this case, mammals appear to be the ones showing a great dependency of codon bias on surrounding base composition. I have shown evidence that suggest that

codon bias is partly dependent on expression levels, but background base composition remains the main determining factor. This has usually been interpreted as evidence for a greater influence of mutation patterns and relaxed selection on mammalian genes. The analyses performed in chapters 4 and 5 show that expression levels of genes are related to regional characteristics, in particular to base composition. G+C content is associated with open chromatin and inner location in the nucleus and it is elevated in highly expressed genes. This could suggest that an elevated G+C composition is important for higher transcription. This being so, third site base composition might not be reflecting mutation patterns only, but rather selection for proper DNA structure suitable for transcription (Vinogradov 2001a; Vinogradov 2003). Were this the case, then relaxed constraints may not be the full explanation for the differences, but rather a greater relevance of chromatin structure for gene expression in larger genomes.

If we turn to gene length distributions, we also find marked differences in how they relate to expression across species. Introns, for example, appear to be under pressure to be reduced in humans as suggested by highly expressed genes having shorter introns than lowly expressed genes. The same appears to be true for the nematode *C. elegans* (Castillo-Davis et al. 2002). Assuming that codon bias is an indicator of expression levels, many other multicellular species should show a similar pattern. However, the same relation is not found in unicellular species such as yeast. Here it is the most highly expressed genes that are the ones with larger introns (Vinogradov 2001b; Vinogradov 2001c). A similar shift in distribution is found when looking at coding region length (Moriyama and Powell 1998).

If reducing gene sizes to minimize transcription and translation costs is favoured by selection, how can we explain the fact that some species show a negative correlation with expression estimates while others show a positive one? A possible explanation is that genes

of different species are shaped by different processes. Alternatively, the general processes influencing sequence evolution are similar across lineages but their relative importance to gene evolution varies. From the second alternative, it could be that reduction of transcription/translation costs may be relevant for all species, including unicellular species such as yeast; in some of them, however, pressure on genome compaction is so great that almost all non-relevant DNA has already been eliminated; any further compaction is likely to interfere with the proper function and regulation of a gene. Therefore selection may favour highly expressed genes to keep their regulatory regions and splicing sites while genes under lower selection may be subject to more deletion events. In contrast, in organisms with larger genomes, where genes have been bombarded with repetitive elements and transposons in their intervening and even coding sequences (Nekrutenko and Li 2001), selection would favour highly expressed genes to reduce length to maximise protein synthesis speed.

What about the relationship between expression and gene location? What are the expectations for smaller genomes in terms of the relationship between gene expression and gene location? The results of the analyses in human genes support the notion that genomic location is strongly determined by transcription competency of chromosomal regions. Regional effects, at least as far as base composition in synonymous sites are concerned, seem to be of less importance in smaller genomes. Following this line of thinking, we may expect that gene order in these genomes would be less determined by the transcription potential of different regions. In these species, co-regulation and operon-like structures might be greater determiners of gene order. Analysis of the *C. elegans* genome has showed that gene order is not random. While co-expression of linked genes is mostly due to operon structures (Lercher et al. 2003), essential genes tend to be found in clusters associated with

lower recombination rates (Pal and Hurst 2003). Interestingly, a recent study analysing Drosophila genes has shown that clusters of coexpressed genes are found along the genome (Spellman and Rubin 2002). In contrast to the results obtained with human genes, in drosophila no evidence for a relation of these clusters with chromosome structural parameters was recovered. Further studies are required to understand the nature of the variations of the relative importance of different factors influencing gene order.

In conclusion, highly/broadly expressed human genes are distinct from the rest of the genes both in terms of their sequence parameters and their location along the chromosomes. Highly expressed genes encode shorter proteins, have shorter introns, higher codon bias and favour particular amino acids in their composition. All of these signatures suggest that, contrary to previous expectations, gene sequence parameters in human genes are tuned for expression efficiency. The results relating expression patterns with gene order and chromosome structure make evident that the human genome despite its size is a highly ordered structure. Results here presented suggest that maximising expression efficiency is a significant factor that has shaped genes and their place in the genome during evolution.

Bibliography

Akashi, H. and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. USA* **99:** 3695-3700.

Balazsi, G., K.A. Kay, A.L. Barabasi, and Z.N. Oltvai. 2003. Spurious spatial periodicity of co-expression in microarray data due to printing design. *Nucleic Acids Res* **31:** 4425-4433.

Castillo-Davis, C.I., S.L. Mekhedov, D.L. Hartl, E.V. Koonin, and F.A. Kondrashov. 2002. Selection for short introns in highly expressed genes. *Nat Genet* **31:** 415-418.

Clay, O., S. Caccio, S. Zoubak, D. Mouchiroud, and G. Bernardi. 1996. Human coding and noncoding DNA: Compositional correlations. *Mol. Phylogenet. Evol.* **5:** 2-12.

Comeron, J.M. and M. Kreitman. 2000. The correlation between intron length and recombination in drosophila: Dynamic equilibrium between mutational and selective forces. *Genetics* **156:** 1175-1190.

Cremer, T. and C. Cremer. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2:** 292-301.

Detours, V., J.E. Dumont, H. Bersini, and C. Maenhaut. 2003. Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett* **546:** 98-102.

Dufton, M.J. 1997. Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? *Journal of Theoretical Biology* **187:** 165-173.

Duret, L. and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. *Proc. Natl Acad. Sci. U.S.A.* **96:** 4482-4487.

Ferreira, J., G. Paolella, C. Ramos, and A.I. Lamond. 1997. Spatial organization of large-scale chromatin domains in the nucleus: A magnified view of single chromosome territories. *J. Cell Biol.* **139:** 1597-1610.

Gouy, M. and C. Gautier. 1982. Codon usage in bacteria - correlation with gene expressivity. *Nucleic Acids Res* **10:** 7055-7074.

Green, P., B. Ewing, W. Miller, P.J. Thomas, and E.D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33:** 514-517.

Hickey, D.A. and B. Benkel. 1986. Introns as relict retrotransposons: implications for the evolutionary origin of eukaryotic mRNA splicing mechanisms. *J. theor. Biol.* **121:** 283-291.

Ishii, M., S. Hashimoto, S. Tsutsumi, Y. Wada, K. Matsushima, T. Kodama, and H. Aburatani. 2000. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68:** 136-143.

Lercher, M.J., T. Blumenthal, and L.D. Hurst. 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Res* **13:** 238-243.

Lynch, M. and A. Kewalramani. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol Biol Evol* **20:** 563-571.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet* **73:** 688-692.

Moriyama, E.N. and J.R. Powell. 1998. Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. *Nucleic Acids Res* **26:** 3188-3193.

Mouchiroud, D., G. Donofrio, B. Aissani, G. Macaya, C. Gautier, and G. Bernardi. 1991. The Distribution of Genes in the Human Genome. *Gene* **100:** 181-187.

Nekrutenko, A. and W.H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17:** 619-621.

Novembre, J.A. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19:** 1390-1394.

Pal, C. and L.D. Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33:** 392-395.

Saccone, S., A. Desario, J. Wiegant, A.K. Raap, G. Dellavalle, and G. Bernardi. 1993. Correlations between Isochores and Chromosomal Bands in the Human Genome. *Proc. Natl Acad. Sci. U.S.A.* **90:** 11929-11933.

Sharp, P.M., T.M.F. Tuohy, and K.R. Mosurski. 1986. Codon usage in yeast - cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14:** 5125-5143.

Spellman, P.T. and G.M. Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* **1:** 5.

Stenico, M., A.T. Lloyd, and P.M. Sharp. 1994. Codon usage in caenorhabditis-elegans - delineation of translational selection and mutational biases. *Nucleic Acids Res* **22:** 2437-2446.

Urrutia, A.O. and L.D. Hurst. 2003. The signature of selection mediated by expression on human genes. *Genome Res* **13:** 2260-2264.

Versteeg, R., B.D. Van Schaik, M.F. Van Batenburg, M. Roos, R. Monajemi, H. Caron, H.J. Bussemaker, and A.H. Van Kampen. 2003. The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes. *Genome Res* **12:** 12.

Vinogradov, A.E. 2001a. Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* **18:** 2195-2200.

---. 2001b. Intron length and codon usage. *J. Mol. Evol.* **52:** 2-5.

---. 2001c. Intron length and codon usage (vol 52, pg 2, 2001). *J. Mol. Evol.* **52:** 310-310.

---. 2003. DNA helix: the importance of being GC-rich. *Nucleic Acids Res* **31:** 1838-1844.

Appendices

Human X Chromosome is Enriched for Male-Specific but not Female-Specific Genes

A Short Note on Gene Order in the Human Genome [spanish]

Selection on Termination Codons in Human Genes

Appendix One

Human X Chromosome is Enriched for Male-Specific but not Female-Specific Genes

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2003. Evidence That the Human X Chromosome Is Enriched for Male-Specific but not Female-Specific Genes. Mol Biol Evol 20: 1113-6.

Evidence That the Human X Chromosome Is Enriched for Male-Specific but not Female-Specific Genes

Martin J. Lercher, Araxi O. Urrutia, and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

There is increasing evidence that X chromosomes have an unusual complement of genes, especially genes that have sex-specific expression. However, whereas in worm and fly the X chromosome has a dearth of male-specific genes, in mice genes that are uniquely expressed in spermatogonia are especially abundant on the X chromosome. Is this latter enrichment true for nongermline, male-specific genes in mammals, and is it found also for female-specific genes? Here, using SAGE data, we show (1) that tissue-specific genes tend to be more abundant on the human X chromosome, (2) that, controlling for this effect, genes expressed exclusively in prostate are enriched on the human X chromosome, and (3) that genes expressed exclusively in mammary gland and ovary are not so enriched. This we propose is consistent with Rice's model of the evolution of sexually antagonistic alleles.

Introduction

Increasing evidence suggests that X chromosomes in diverse species contain unusual complements of genes, especially sex-specific genes. In *Caenorhabditis elegans*, sperm-enriched and germline-intrinsic genes are nearly absent from the X chromosome (Reinke et al. 2000). Similarly, in Drosophila, there is a dearth of male-specific accessory gland protein genes on the X chromosome (Swanson et al. 2001). More generally, Drosophila's testes-specific genes tend to be especially abundant on autosomes, having been derived by retroposition from X-linked genes (Betran, Thornton, and Long 2002). This observation may be explained by natural selection favoring those new retrogenes that moved to autosomes and avoided the spermatogenesis X inactivation (Betran, Thornton, and Long 2002; Boutanaev et al. 2002). This is supported by the finding that clusters of testes-specific genes are described in the only known segment of the X chromosome devoid of the MSL-induced H4 acetylation (Boutanaev et al. 2002). The same may also apply in *C. elegans*, it too having an inactive X chromosome in the male germline (Fong et al. 2002; Kelly et al. 2002; Reuben and Lin 2002). Some credence is given to this hypothesis from the finding that in worm, the X chromosomes in the XX germline are silenced only in early meiosis (Kelly et al. 2002) and that ovary-expressed genes are present on the X chromosome (Reinke et al. 2000).

Is germline X chromosome inactivation (or more generally male-specific X chromosome–associated chromatin remodeling complexes [Boutanaev et al. 2002]) the sole cause of the unusual gene complement of X chromosomes? In contrast to the above, human genes whose mutants disrupt sexual development are especially common on the X chromosome (Saifi and Chandra 1999). Similarly, Wang et al. (2001), using a cDNA subtraction method, identified 25 mouse genes that appeared to be uniquely expressed in spermatogonia: three of these were Y linked and 10 were X linked. Were gene distribution random, they argued that about an order of magnitude fewer X-linked genes would be expected.

Rice's Hypothesis

One interpretation (Hurst 2001; Wang et al. 2001) of this enrichment of spermatogonial genes on the mammalian X chromosome is that it is a consequence of the evolution of sexually antagonistic alleles (i.e., alleles that are beneficial to one sex but detrimental to the other). Rice (1984) noted that, despite the fact that an X chromosome spends only one third of its time in the male germline, a perfectly recessive allele of an X-linked gene that is favorable to the hemizygous sex (hereafter males) is much more likely to spread than an autosomal counterpart. This is because selection would act strongly on the hemizygously expressed favorable effects, whereas the deleterious effects in females would initially be masked, owing to heterozygosity in females. The autosomal counterpart would have all effects hidden and hence be likely to be lost.

If the allele is not perfectly recessive then for the autosomal case, the beneficial effects in males must counterbalance the deleterious effects in females. For the X-linked gene the beneficial effects could be relatively weak if the allele has no great fitness consequences in heterozygous females. Hence, even an allele with great negative fitness consequences when homozygous in females might spread. Consequentially, once the allele attains a significant frequency, the evolution of modifiers that force the gene to be expressed only in males is expected (Rice 1984). As most mutations are recessive, we expect an enrichment of male-specific genes on the X chromosome. Comparable logic predicts enrichment of male-benefit traits on the Y chromosome as well.

Support for the premise of Rice's model comes from the findings that the X chromosome appears to harbor a disproportionately large amount of variation in sexually selected traits (Reinhold 1998) and is, more generally, enriched for sexually antagonistic fitness variation (Gibson, Chippindale, and Rice 2002). These findings need not, however, reflect a greater abundance of genes of any given type on the X chromosome.

If Rice's hypothesis holds, we might make two predictions. First, genes expressed exclusively in other male-specific tissues will also be especially common on

Key words: X chromosome, prostate, mammary gland, ovary, sexual antagonism, gene location.

E-mail: l.d.hurst@bath.ac.uk.

Mol. Biol. Evol. 20(7):1113–1116. 2003
DOI: 10.1093/molbev/msg131
Molecular Biology and Evolution, Vol. 20, No. 7,

the mammalian X chromosome, assuming there is no interaction with inactivation of the X chromosome (the X chromosome in murine spermatogenesis is inactivated probably by a highly conserved mechanism [Reuben and Lin 2002]). We examine this issue by looking at genes that are expressed exclusively in a somatic male-specific tissue, the prostate. Second, genes expressed in female-specific tissues need not be enriched on the X chromosome.

The latter is owing to the fact that, under Rice's model, two forces act antagonistically. Consider first a dominant allele that is beneficial to females but detrimental to males. As the X chromosome spends two thirds of its time in females, the favorable effects of the allele are evident more commonly than the deleterious effects in males, compared with the same dominant allele when autosomal. This acts as a force to increase the chances that a female-benefit /male-detriment allele might spread, were it X linked, and hence is a force leading to enrichment on the X chromosome of female-specific genes (after a modifier has suppressed the genes' expression in males). However, this force will be counterbalanced by the greater relative ease of female-advantageous/male-detrimental alleles to spread on autosomes when partially recessive, the X-linked version being relatively heavily counter selected from the outset owing to hemizygosity in males. Hence enrichment of female-specific genes on the X chromosome is not necessarily expected. We shall examine this issue by investigating the genomic location of genes expressed exclusively in human mammary gland or ovary.

Tissue Specificity and the Human X Chromosome

One important difference between the present analysis and all prior analyses is that we control for tissue specificity. We recently showed that on the average, genes on the X chromosome are expressed in fewer tissues than genes on autosomes (Lercher, Urrutia, and Hurst 2002). One might speculate that this may be the result of selection to minimize the deleterious effects of mutations in X-linked genes. This speculation aside, if X-linked genes do tend to be tissue specific per se, then we expect enrichment on the X chromosome for any class of genes that are tissue specific regardless of sex specificity. This could indeed go some way to explain prior results. Hence, we establish a data set of expression patterns for over 8,000 genes but then extract only those expressed in just one tissue.

Materials and Methods

The SAGE Data Set

We used publicly available data from Serial Analysis of Gene Expression (Velculescu et al. 1995; SAGE). From SAGEmap (Lash et al. 2000) at NCBI, we obtained a reliable mapping of UniGene (Schuler et al. 1996) groups to *Nla*III SAGE tags. Each UniGene group consists of all GenBank sequences representing the same human gene. In the remainder, we will refer to each such group as a gene and represent it by its longest RefSeq sequence. Tags mapping to more than one gene were excluded. We located 11,612 RefSeq genes on the August 2001 Golden Path assembly of the human genome (http://genome.cse.ucsc.edu/), each labeled unambiguously by at least one SAGE tag. This set of gene/tag combinations was cross-linked to the quantitative expression profiles at SAGEmap. Positive expression was seen in 8,367 genes in at least one out of 35 libraries representing 14 normal (i.e., non-pathological) tissues. If a tag had been counted only once in one tissue, this was most likely due to a sequencing error, and we discounted the observation. Adding all counts for libraries representing the same tissue type, we then calculated breadth of expression (number of tissues with positive expression) for each gene. Genes were counted as tissue specific if they were expressed in only one of the 14 tissues.

Statistics

To determine the significance of the observed number of genes of a given class (prostate, ovary/mammary) on the X chromosome against null expectations, we employed a randomization strategy. We reassigned all genes at random to chromosomes while maintaining the total gene count, the total count of genes within each class, and the total number of genes on each chromosome as found in the original data set. The P value was then specified as the proportion of randomizations in which the actual number, or a greater number, of genes within the class in question appeared on the X chromosome.

The expectations for the number of genes on the X chromosome can be derived by this method or by partitioning the data into tissue-specific genes that are not sex specific and using the X:A ratio to deduce the expected number of X-linked genes within any given class, given the total number of genes in this class. Both method estimates are provided. The first estimate given below is always from the X:A ratio, and the second is from randomization.

Results

Our prior work suggested that genes on the X chromosome are not expressed in as many tissues as autosomal genes (Lercher, Urrutia, and Hurst 2002). Does it follow that the X chromosome has more tissue-specific genes? If we examine genes expressed in at least nine of the 14 tissues (N = 1,897) (our prior definition of housekeeping genes [Lercher, Urrutia and Hurst 2002]), we find 50 that are X linked (i.e., 2.7% of the total). By contrast, of genes expressed in three or fewer tissues (N = 3,441), 3.8% are X linked ($P < 0.02$ by randomization, two tailed). Of those expressed in just one tissue, 3.6% of the total of 1,511 are X linked. Although this latter result is not significant at the 5% level ($P = 0.069$, by randomization, two tailed), given the apparent tendency, it is best to be conservative and to control for tissue specificity.

Are prostate-specific genes especially prevalent on the X chromosome? Of the tissue-specific genes that are not expressed in the sex-specific tissues (ovary, mammary gland, or prostate)1,046 are autosomal and 35 (3.3%) are X linked. Of the prostate-specific genes, 189 are autosomal compared with 13 (6.9%) that are X linked. This represents an approximate doubling of the frequency of

prostate-specific genes on the X chromosome and represents a significant enrichment (6.5/ 7.3 are expected, $P = 0.02$, one tailed, derived by 100,000 randomizations). Pairwise Blast of all of the X-linked prostate-specific genes against all the others on the X chromosome revealed no duplicate genes, so the enrichment is not owing to higher rates of duplication on the X chromosome.

It may be notable that our estimate of the extent of the enrichment of male-specific genes (an approximate doubling) is lower than that observed by Wang et al. (2001). This is unlikely to be owing solely to methodological differences (of which control for tissue specificity would be one), as the difference appears to be quite large: Wang et al. report that nearly 40% of the spermatogonia-specific genes are X linked, which compares with just 7% for prostate. Perhaps there is significant heterogeneity between male-specific tissues? When high-quality expression data is available for more male-specific tissues, this should be testable.

In our sample, female-specific genes, in contrast to the male-specific genes, show no X-linked enrichment when compared against tissue-specific genes. Whereas 222 genes expressed in ovary or mammary gland are autosomal, only six (2.7%) are X-linked genes expressed in either tissue. If anything then, female-specific genes are underrepresented on the X chromosome, although the difference is not statistically significant (six observed, 7.4/8.2 are expected, $P = 0.33$). Analyzing ovary alone (under the supposition that some mammary gland genes might also be in male breast tissue) does not alter the conclusions: 107 are autosomal, four are X linked, and four are expected (by both methods) ($P = 0.57$).

Discussion

The above results provide support, by no means definitive, that Rice's hypothesis may be important to understanding mammalian X chromosome evolution. However, this should be regarded as a provisional interpretation, as numerous caveats must be noted. For example, in several years time SAGE data will, no doubt, be available for many more tissues, in which case, it is all but inevitable that some of our "tissue-specific" genes will turn out not to be tissue specific at all, just expressed in relatively few tissues. This need not prove be too problematic for the current provisional interpretation, as Rice's model does not require the genes to be expressed exclusively in one tissue. However, more problematically, it may yet prove to be the case that some "ovary-specific" genes are in fact germline-specific genes and expressed in both males and females. Prior evidence suggests that genes expressed in both germlines are not enriched on the X chromosome (Wang et al. 2001). SAGE analysis on testicular tissue would allow us to eliminate this possibility.

Further, in our presentation of Rice's hypothesis, we assumed the presence of alleles expressed in both sexes for genes already present on the X chromosome. It is uncertain whether it is reasonable to suppose that there were genes expressed both in prostate and in females as well. Similarly, it may possibly be that the genes were originally autosomal and their sexually antagonistic phenotype predisposed them to becoming X linked (Charlesworth and Charlesworth 1980). Even were our finding statistically robust, the interpretation is by no means certain.

Despite the above caveats, given the present results and those of Wang et al. (2001), we can tentatively suggests that, consistent with Rice's hypothesis, the mammalian X chromosome is enriched for male-specific but not female-specific genes. What also of the Y chromosome? As expected, in our sample, no mammary-specific or ovary-specific genes are Y linked. Two of the seven Y-linked sequences in our sample were prostate specific, the others being expressed (apparently in a sex-specific manner) either in brain or in peritoneum. Overall enrichment of prostate-specific genes on the X or Y chromosome is significant ($P = 0.01$, by randomization).

The description of some brain-specific, Y-linked genes is especially notable, as it has also recently been suggested that selection for sex differences in cognitive ability might explain why genes that affect cognitive ability appear also to be enriched on the X chromosome (Zechner et al. 2001). Although there are too few brain-specific, Y-linked genes to perform meaningful statistics, there may be weak enrichment of these: we expect about one and observe three. This and the putative X chromosome enrichment may also reflect the processes envisaged by Rice. However, brain-specific genes (white matter, astrocyte, and thalamus) in our sample are not enriched on the X chromosome: we expect 13.5/12.9 X-linked genes, which compares with 14 observed ($P = 0.43$) (of 406 brain-specific genes, 389 are autosomal and 14 [2.1%] are X linked; of non–brain-specific, non–sex-specific genes, 657 are autosomal and 21 [3.2%] are X linked). This brain sample presumably includes both sex-specific and non–sex-specific genes, and it would be valuable to return to the issue using direct expression assays when sex specificity of gene expression in non–sex-specific tissues can be assayed.

Acknowledgments

We wish to thank two anonymous referees for comments on an earlier version of the manuscript. M.J.L. is funded by the Wellcome Trust, A.O.U. is funded by an Overseas Research Students award and a CONACyT grant, and L.D.H. is funded by the UK Biotechnology and Biosciences Research Council.

Literature Cited

Betran, E., K. Thornton, and M. Long. 2002. Retroposed new genes out of the X in Drosophila. Genome Res. **12**:1854–1859.

Boutanaev, A. M., A. I. Kalmykova, Y. Y. Shevelyou, and D. I. Nurminsky. 2002. Large clusters of co-expressed genes in the Drosophila genome. Nature **420**:666–669.

Charlesworth, D., and B. Charlesworth. 1980. Sex-differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. Genet. Res. **35**:205–214.

Fong, Y. Y., L. Bender, W. C. Wang, and S. Strome. 2002. Regulation of the different chromatin states of autosomes and

X chromosomes in the germ line of *C. elegans*. Science **296**:2235–2238.

Gibson, J. R., A. K. Chippindale, and W. R. Rice. 2002. The X chromosome is a hot spot for sexually antagonistic fitness variation. Proc. R. Soc. Lond. B Biol. Sci. **269**:499–505.

Hurst, L. D. 2001. Evolutionary genomics—sex and the X. Nature **411**:149–150.

Kelly, W. G., C. E. Schaner, A. F. Dernburg, M. H. Lee, S. K. Kim, A. M. Villeneuve, and V. Reinke. 2002. X-chromosome silencing in the germline of *C. elegans*. Development **129**:479–492.

Lash, A. E., C. M. Tolstoshev, L. Wagner, G. D. Schuler, R. L. Strausberg, G. J. Riggins, and S. F. Altschul. 2000. SAGEmap: a public gene expression resource. Genome Res. **10**:1051–1060.

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31**:180–183.

Reinhold, K. 1998. Sex linkage among genes controlling sexually selected traits. Behav. Ecol. Sociobiol. **44**:1–7.

Reinke, V., H. E. Smith, J. Nance et al. (11 co-authors). 2000. A global profile of germline gene expression in *C. elegans*. Mol. Cell. **6**:605–616.

Reuben, M., and R. Lin. 2002. Germline X chromosomes exhibit contrasting patterns of histone H3 methylation in *Caenorhabditis elegans*. Dev. Biol. **245**:71–82.

Rice, W. R. 1984. Sex-chromosomes and the evolution of sexual dimorphism. Evolution **38**:735–742.

Saifi, G. M., and H. S. Chandra. 1999. An apparent excess of sex- and reproduction-related genes on the human X chromosome. Proc. R. Soc. Lond. B Biol. Sci. **266**:203–209.

Schuler, G. D., M. S. Boguski, E. A. Stewart et al. (101 co-authors). 1996. A gene map of the human genome. Science **274**:540–546.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98**:7375–7379.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. Serial analysis of gene expression. Science **270**:484–487.

Wang, P. J., J. R. McCarrey, F. Yang, and D. C. Page. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat. Genet. **27**:422–426.

Zechner, U., M. Wilda, H. Kehrer-Sawatzki, W. Vogel, R. Fundele, and H. Hameister. 2001. A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? Trends Genet. **17**:697–701.

William Jeffery, Associate Editor

Accepted March 17, 2003

Appendix Two

A Short Note on Gene Order in the Human Genome

Un modelo de orden génico en el genoma humano

Araxi Urrutia Odabachian

Biology & Biochemistry, University of Bath, BA2 7AY, UK. bspauo@bath.ac.uk.

El **genoma humano** contiene aproximadamente 32000 **genes** que codifican **proteínas** que regulan el desarrollo, metabolismo y demás funciones en el organismo. Todos los genes del núcleo están distribuidos entre los 23 **cromosomas** que constituyen el genoma humano. La totalidad de los genes ocupan menos del 5% del genoma, por lo que de estar distribuidos de manera uniforme cada gen debería estar aislado de los demás. Este, sin embargo, no es el caso, la densidad de genes varía significativamente a lo largo de los cromosomas. Grandes regiones están prácticamente desiertas, mientras en otras los genes están tan próximos que se sobreponen. ¿Qué determina la posición y el orden de los genes en el genoma?

Normalmente se asume que las presiones selectivas sobre la distribución de genes son mínimas. De ser así, la posición y orden actual de los genes serían entonces las que se esperan por azar. Sin embargo, los análisis realizados han revelado patrones, difíciles de explicar por simple azar, que sugieren que la distribución de genes está ligada a la compleja estructura de los cromosomas.

Los cromosomas se encuentran divididos en regiones que difieren tanto en estructura como en composición de bases. Así, de las cuatro bases nitrogenadas en el DNA, la proporción de Guanina y Citosina respecto a Adenina y Timina puede ser tan alta como 60% en algunas regiones del genoma, y en otras tan baja como 35%. La densidad de genes está asociada a estas variaciones en composición de nucleótidos. Se ha observado que los genes tienden a concentrarse en regiones ricas en Guanina y Citosina (G+C).

De modo independiente, desde finales de 1800's, usando técnicas de coloración Giemsa, en células en **mitosis,** se identificaron bandas transversales a lo largo de los cromosomas que varían en la intensidad de la coloración. Ferreira, Carmo-Fonseca y colaboradores de la Universidad de Lisboa en Portugal observaron que, durante la división celular, las bandas más pálidas se replican antes que las bandas más obscuras. Las bandas más obscuras son más compactas y tienden a ocupar regiones hacia la periferia del núcleo. Se ha observado que las variaciones en bases nitrogenadas están relacionadas con las bandas de Giemsa: las bandas más claras y menos compactas contienen una mayor proporción de G+C, mientras que las bandas más obscuras y compactas son más ricas en Adenina y Timina. Podría esperarse por tanto que la densidad de genes siga de cerca los patrones de coloración de los cromosomas. En efecto, aproximadamente 80% de los genes se concentran en las bandas mas claras.

Podemos concluir entonces que en el genoma hay regiones más habitables que otras. Aún podría ser que dentro de las regiones competentes los genes podrían estar ordenados al azar. De nuevo, los datos obtenidos sugieren lo contrario. El orden de los genes depende de sus patrones de actividad. Los genes varían en sus niveles de actividad ya que no todas las proteínas son requeridas en todos los **tipos celulares** y/o en las mismas cantidades. En una muestra de 10000 genes, Lercher y colaboradores de la Universidad de Bath en el Reino Unido, observaron que los genes se agrupan según el número de tejidos en los que se expresan, esto es; genes que se expresan en numerosos tejidos y genes que se expresan en pocos tejidos tienden a estar cerca de los genes de expresión similar.

El modelo de *compartamentalización funcional* del genoma provee un marco común para explicar las irregularidades tanto en la densidad como en el orden de los genes a lo largo de los cromosomas. Este modelo propone que la estructura en una región cromosómica

particular es un factor determinante para el potencial de actividad de los genes. Las regiones correspondientes a bandas de Giemsa más claras y menos compactas se sitúan en la periferia del núcleo donde los complejos de transcripción pueden acceder más fácilmente. Es ahí donde se espera la mayor densidad de genes tal y como se ha observado. Cremer y Cremer, basados en las universidades de Alemania Ludwig Maximilians y Ruprecht Karls respectivamente, establecieron la hipótesis de que aquellos genes que se expresan en más tipos celulares, deberían estar concentrados donde la actividad de transcripción es más intensa. Los datos de Lercher y colaboradores son consistentes con este escenario. De particular relevancia es la observación de que los genes de mayor espectro de expresión, están en regiones ricas en G+C. Si el modelo propuesto por Cremer & Cremer es correcto, entonces podemos esperar que los genes de amplia expresión se localicen en las bandas de Giemsa más claras y menos compactas de los cromosomas.

De confirmarse esto último, la visión del genoma como depósito pasivo de genes tendría que dar paso a un modelo en el que la estructura de los cromosomas juega un papel activo en la determinación de los patrones generales de expresión y en la evolución de las especies. Por ejemplo, el efecto de mutaciones en las que secciones de los cromosomas son eliminadas o duplicadas podría derivarse no solo de la ausencia de los genes eliminados o añadidos sino de los cambios en la organización de los cromosomas dentro del núcleo que ocurren como consecuencia de esta mutación.

Las observaciones anteriores sugieren que factores relacionados con la estructura cromosómica son un factor relevante en la determinación del orden de los genes en **eucariontes**, en particular en genomas mamíferos. En contraste, en los genomas de organismos **procariontes,** como las bacterias, es común que genes implicados en un mismo proceso celular se encuentren agrupados en **operones** en el genoma y su actividad es

regulada en conjunto. En los organismos eucariontes, como los insectos y los mamíferos hay poca evidencia de que la co-regulación sea un determinante importante en la localización para la mayoría de los genes.

Agradecimientos. A Humberto Gutiérrez y Yazmín Odabachian por sus valiosas contribuciones a este manuscrito. A CONACYT, ORS y Korner Award por financiamiento.

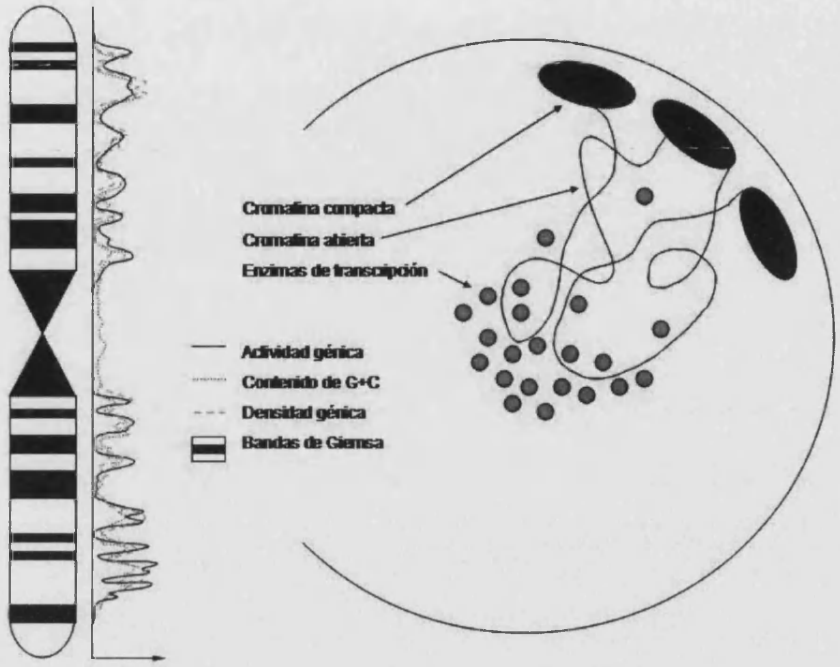


Figura1. La figura muestra los puntos más relevantes del modelo propuesto de orden génico y su relación con la estructura cromosómica. El esquema tiene por propósito mostrar las expectativas del modelo, sin embargo, algunos de los puntos no han sido confirmados. Los niveles de actividad promedio de los genes en cada región correlaciona con la densidad génica, el contenido de guanina y citosina y las bandas de Giemsa. Las bandas de Giemsa a su vez se distribuyen desigualmente dentro del núcleo. Las bandas más obscuras y compactas se sitúan en la periferia, mientras que las bandas más claras y ricas en genes se encuentran en el interior del núcleo donde el acceso a las enzimas de transcripción es mayor.

Appendix Three

Selection on Termination Codons in Human Genes

SELECTION ON TERMINATION CODON USAGE IN HUMAN GENES

Araxi Urrutia Odabachian

ABSTRACT

I analyse the use of termination codons in a sample of 2396 human genes and its relation with GC content, codon bias, expression breadth and rates of synonymous substitutions. We found that TGA is the termination codon used in half of the sample. The strength of the preference is influenced by GC content, breadth of expression, and to a lesser extent, codon usage bias. We conclude that, in contrast with other species, the choice of termination codon in human genes is mainly driven by nucleotide content and dinucleotide concentrations.

Codon usage bias has been extensively studied in several unicellular and multicellular species (Marais and Duret 2001), and it has been found that the level of codon bias correlates with expression levels, highly expressed genes showing higher codon bias (Gouy and Gautier 1982; Sharp et al. 1986; Stenico et al. 1994), and it is inversely correlated with rates of synonymous substitutions (Powell and Moriyama 1997). However the factors determining the use of termination codons has received less attention in part because of their low frequency. In unicellular and invertebrate species the choice of stop codon has been found to be not random but show biases (Duret and Mouchiroud 1999; Sharp and Bulmer 1988). The choice of stop codon (CSC) is related with levels of codon usage bias (Duret and Mouchiroud 1999; Sharp and Bulmer 1988). In mammals because of the great variation in nucleotide content across different regions of the genome (Bernardi 1995; Bernardi et al. 1997), studies so far have failed to determine if there are

any significant selective pressures determining codon usage bias (Eyre-Walker 1994; Eyrewalker 1991; Urrutia and Hurst 2003). Here we analyse the choice of stop codons in human genes and its relationship with nucleotide content within the coding sequence. Also we study the relationship between CSC and variables that have been shown to be important in other species such as codon usage bias, expression patterns and rates of evolution.

MATERIALS AND METHODS

Coding sequences from 2396 human genes were used for this study. Termination codon (SC) was determined for each sequence. Information about breadth of expression and synonymous rates of substitution were obtained from Duret and Mouchiroud database (2000). G+C content at third sites and termination codon used was obtained for each sequence and codon usage bias was measured by MCB method correcting for background nucleotide biases (Urrutia and Hurst 2003).

RESULTS AND DISCUSSION

We investigated the distribution of stop codons in our sample. In figure 1, it can be observed that there is a strong preference in the sample towards TGA as a termination codon over TAA and TAG codons. This is probably related to the avoidance of TpA dinucleotides that is observed in the human genome even at non-coding regions (Karlin and Mrazek 1996). However we found little effect of the levels of dinucleotide bias and CSC (Fig. 2a).

We investigated whether the termination codon preferences are related to characteristics of the coding sequence such as nucleotide content and breadth of expression. We found that CSC is particularly influenced by the GC content at third sites of the sequence, there being an increase in the use of TGA and TAG codons in genes with higher GC content, while the use of TAA is greatly reduced (Fig. 2b). We also observe an increase in preference of TAA in broadly expressed genes (Fig. 2c). A much weaker interaction is observed between codon usage bias corrected by background nucleotide biases and the choice of termination codon, there is a slight increase in the use of TAG at the expense of TAA (Fig. 2d). A similarly weak effect on termination codon by gene length is observed in this case TAA is avoided in larger genes (Fig. 2e).

Here we showed that the choice of termination codons in human genes is not random but that there are strong biases in the use of stop codons. The most important factor determining CSC appears to be the avoidande of the dinucleotide TpA. Almost half of the genes in the sample use TGA as termination codon. However only a weak effect can be observed between the extent of dinucleotide biases on termination codon choice. The two factors that are strongly related with termination codon choice are GC3 content and expression breadth, while the first indicates that mutation processes or biased gene conversion favouring specific nucleotides are in action, the second could be indicative of selective pressures favouring a specific choice in termination codon. The fact that these to factors are interrelated makes it difficult to desentangle which is the most prominent variable.

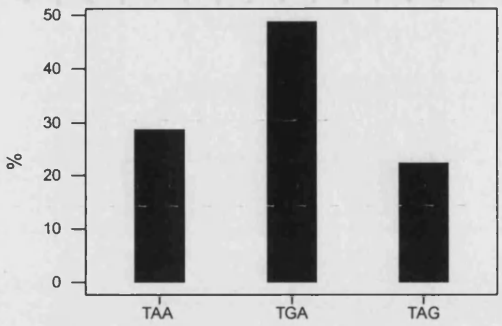
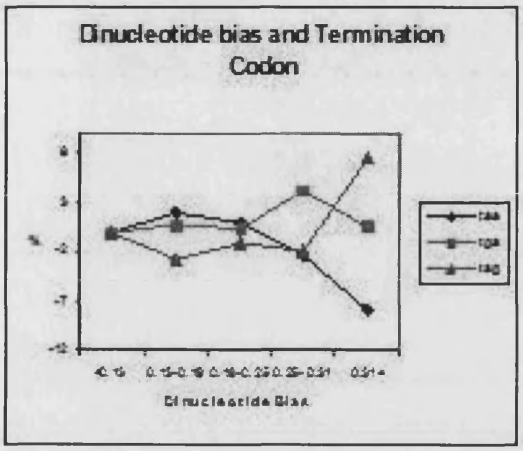
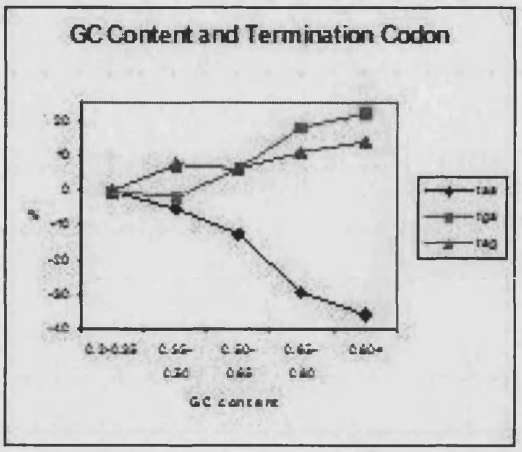
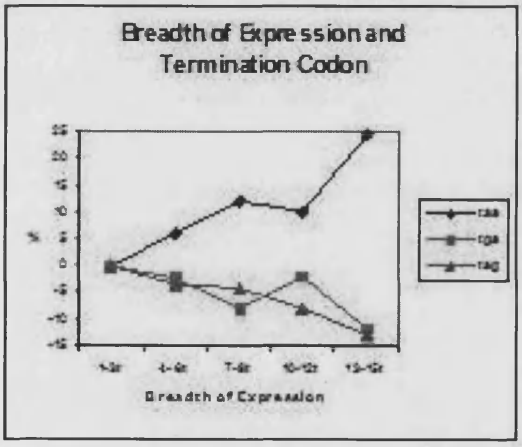
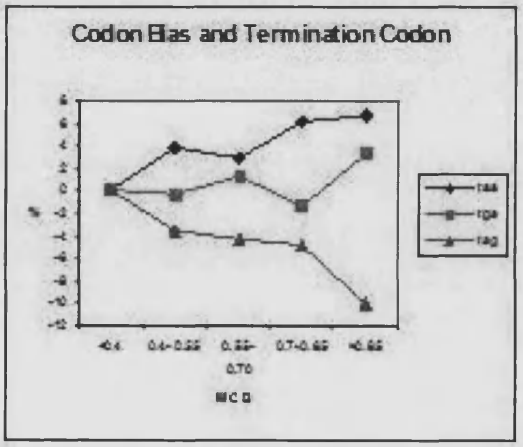


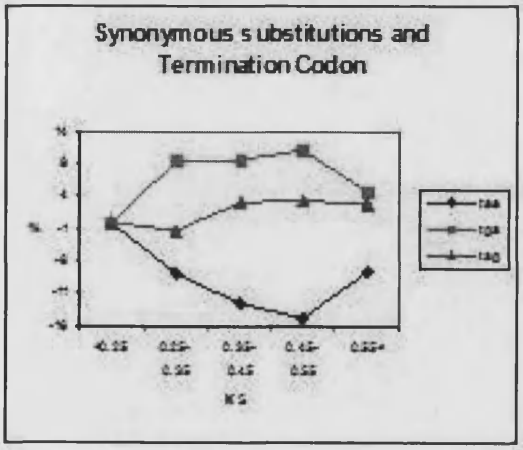
Figure 1. Stop codon usage distribution in sample of 2396 human genes.


Dinucleotide bias and Termination Codon


GC Content and Termination Codon


Breadth of Expression and Termination Codon


Codon Bias and Termination Codon


Synonymous substitutions and Termination Codon

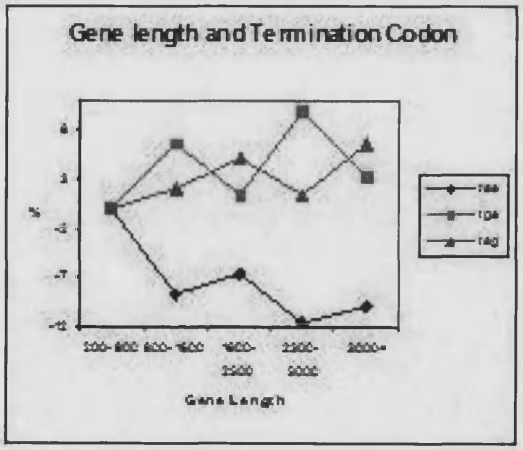

Gene length and Termination Codon

Figure 2. Relationship between choice of termination codon and A. dinucleotide bias, B. G+C content, C. breadth of expression, D. codon usage bias, E. rate of synonymous substitutions, F. gene length. The graphics show the differences in percentage for each termination codon with respect to the value of the lowest values in the x axis.

REFERENCES

Bernardi, G. 1995. The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* **29:** 445-476.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1997. Isochores and synonymous substitutions in mammalian genes. In *DNA and Protein Sequence Analysis* (eds. M.J. Bishop and C.J. Rawlings). IRL Press, Oxford.

Duret, L. and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, Arabidopsis. *Proc. Natl Acad. Sci. U.S.A.* **96:** 4482-4487.

---. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17:** 68-74.

Eyre-Walker, A. 1994. Dna mismatch repair and synonymous codon evolution in mammals. *Mol. Biol. Evol.* **11:** 88-98.

Eyrewalker, A.C. 1991. An Analysis of Codon Usage in Mammals - Selection or Mutation Bias. *J. Mol. Evol.* **33:** 442-449.

Gouy, M. and C. Gautier. 1982. Codon usage in bacteria - correlation with gene expressivity. *Nucleic Acids Res* **10:** 7055-7074.

Karlin, S. and J. Mrazek. 1996. What drives codon choices in human genes? *J. Mol. Biol.* **262:** 459-472.

Marais, G. and L. Duret. 2001. Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. *J. Mol. Evol.* **52:** 275-280.

Powell, J.R. and E.N. Moriyama. 1997. Evolution of codon usage bias in Drosophila. *Proc. Natl Acad. Sci. U.S.A.* **94:** 7784-7790.

Sharp, P.M. and M. Bulmer. 1988. Selective Differences among Translation Termination Codons. *Gene* **63:** 141-145.

Sharp, P.M., T.M.F. Tuohy, and K.R. Mosurski. 1986. Codon usage in yeast - cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14:** 5125-5143.

Stenico, M., A.T. Lloyd, and P.M. Sharp. 1994. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. *Nucleic Acids Res* **22:** 2437-2446.

Urrutia, A.O. and L.D. Hurst. 2003. The signature of selection mediated by expression on human genes. Under review at Genome Research.